



University
of Glasgow

Leelanupab, Teerapong (2012) *A ranking framework and evaluation for diversity-based retrieval.*

PhD thesis

<http://theses.gla.ac.uk/3442/>

Copyright and moral rights for this thesis are retained by the author

A copy can be downloaded for personal non-commercial research or study, without prior permission or charge

This thesis cannot be reproduced or quoted extensively from without first obtaining permission in writing from the Author

The content must not be changed in any way or sold commercially in any format or medium without the formal permission of the Author

When referring to this work, full bibliographic details including the author, title, awarding institution and date of the thesis must be given



University of Glasgow | School of
Computing Science

A Ranking Framework and Evaluation for Diversity-Based Retrieval

Teerapong Leelanupab

A thesis submitted for the degree of
Doctor of Philosophy

School of Computing Science
College of Science and Engineering
University of Glasgow

1 June 2012

Abstract

There has been growing momentum in building information retrieval (IR) systems that consider both *relevance* and *diversity* of retrieved information, which together improve the usefulness of search results as perceived by users. Some users may genuinely require a *set* of multiple results to satisfy their information need as there is no single result that completely fulfils the need. Others may be uncertain about their information need and they may submit ambiguous or broad (faceted) queries, either intentionally or unintentionally. A sensible approach to tackle these problems is to diversify search results to address all possible senses underlying those queries or all possible answers satisfying the information need. In this thesis, we explore three aspects of diversity-based document retrieval: 1) recommender systems, 2) retrieval algorithms, and 3) evaluation measures.

This first goal of this thesis is to provide an understanding of the need for diversity in search results from the users' perspective. We develop an interactive recommender system for the purpose of a user study. Designed to facilitate users engaged in exploratory search, the system is featured with content-based browsing, aspectual interfaces, and diverse recommendations. While the diverse recommendations allow users to discover more and different aspects of a search topic, the aspectual interfaces allow users to manage and structure their own search process and results regarding aspects found during browsing. The recommendation feature mines implicit relevance feedback information extracted from a user's browsing trails and diversifies recommended results with respect to document contents. The result of our user-centred experiment shows that result diversity is needed in realistic retrieval scenarios.

Next, we propose a new ranking framework for promoting diversity in a ranked list. We combine two distinct result diversification patterns; this leads to a general framework that enables the development of a variety of ranking algorithms for diversifying documents. To validate our proposal and to gain more insights into approaches for diversifying documents, we empirically compare our integration framework against a common ranking approach (i.e. the probability ranking principle) as well as several diversity-based ranking strategies. These include maximal marginal relevance, modern portfolio theory, and sub-topic-aware diversification based on sub-topic modelling techniques, e.g. clustering, latent Dirichlet allocation, and probabilistic latent semantic analysis. Our findings show that the two diversification patterns can be employed together to improve the effectiveness of ranking diversification. Furthermore, we find that the effectiveness of our framework mainly depends on the effectiveness of the underlying sub-topic modelling techniques.

Finally, we examine evaluation measures for diversity retrieval. We analytically identify an issue affecting the de-facto standard measure, novelty-biased discounted cumulative gain (α -nDCG). This issue prevents the measure from behaving as desired, i.e. assessing the effectiveness of systems that provide complete coverage of sub-topics by avoiding excessive redundancy. We show that this issue is of importance as it highly affects the evaluation of retrieval systems, specifically by overrating top-ranked systems that repeatedly retrieve redundant information. To overcome this issue, we derive a theoretically sound solution by defining a safe threshold on a query-basis. We examine the impact of arbitrary settings of the α -nDCG parameter. We evaluate the intuitiveness and reliability of α -nDCG when using our proposed setting on both real and synthetic rankings. We demonstrate that the diversity of document rankings can be intuitively measured by employing the safe threshold. Moreover, our proposal does not harm, but instead increases the reliability of the measure in terms of discriminative power, stability, and sensitivity.

Acknowledgements

This thesis has been, in the first place, an extensive journey full of challenges and excitement, despite the alternating anguish and joy it entails. The completion of this thesis would not have been possible without the support and encouragement of my friends, family, and colleagues. Many people contributed to this thesis, directly or indirectly, either by discussing the ideas, reviewing my research work, or providing feedback. To most of them, I owe a special “thank you”.

First and foremost, I am sincerely and heartily grateful to my supervisor, Professor Joemon M. Jose, who patiently steered and supported me over the course of this Ph.D. He deserves much gratitude for having enough faith in my skills to grant me this opportunity and freedom to research. I am also thankful to him for his constant intellectual stimulation, a wonderful working environment, and much support throughout my years at Glasgow. I was very fortunate to be able to study with him in the School of Computing Science at the University of Glasgow.

Special thanks to Professor Keith van Rijsbergen, my other supervisor, for his wisdom, advice and guidance throughout this work. It has been a privilege to have the benefit of his counsel.

I would also like to thank Guido Zuccon for his fresh ideas and knowledge that helped tremendously in shaping my thesis. His tenacity and incredibly positive attitude towards research have been a constant source of inspiration for me. It has been greatly appreciated. He has been of invaluable benefit for discussions and collaboration.

To all my friends and colleagues in the Glasgow IR group: thank you for your support, and for keeping me sane and happy. Special thanks to Phil McParlane, Jesús Rodríguez Pérez, and Stewart Whiting for providing the pre-submission reviews, which bring this thesis into the final form. I am also grateful to all past and present members of my research group: Leif Azzopardi, Frank Hopfgartner, Yue Feng, Álvaro Huertas, Anuj Goyal, Thierry Urruty, and Martin Halvey. Thank you for all your help and providing early feedback on my research.

This work was developed under the funding of the Royal Thai Government, and the European Commission via the K-Space project (Knowledge Space of Technology to Bridge the Semantic Gap). I would like to thank them for their financial support.

I dedicate this thesis to my family, especially my parents and sisters: Pichai and Vipada Leelanupab, and Leena Hatko. I am forever indebted to them who have always encouraged me to settle for nothing but the best. Specially, my father has been a great mentor and advisor to me throughout my life. His support and encouragement played a tremendous role in my decision to pursue a Ph.D.

Last and most of all, thank you, Soontaree Petchdee, for all your sacrifice and support throughout this journey. It is a testament to your strength that we have made it through this Ph.D. together. This dissertation is as much yours as it is mine.

Contents

I	Introduction and Background	3
1	Introduction	4
1.1	Diversity in Information Retrieval	6
1.1.1	Intrinsic Diversity	7
1.1.2	Extrinsic Diversity	7
1.2	Thesis Statement	8
1.2.1	Research Questions	9
1.3	Contributions	10
1.4	Roadmap of the Thesis	11
1.5	Publications	13
2	Retrieval Models, Tasks, and Evaluation	16
2.1	Introduction	16
2.2	Fundamental Concepts of Information Retrieval	16
2.2.1	Boolean Model	17
2.2.2	Vector Space Model	18
2.2.3	Probabilistic Models	20
2.2.3.1	BM25 Model	22
2.3	Information Retrieval Tasks	23
2.3.1	Ad-hoc Document Retrieval	23
2.3.2	Diversity-Based Document Retrieval	24
2.4	Evaluation in Information Retrieval	26
2.4.1	Experimental Methodologies	26
2.4.1.1	System-Oriented Evaluation	27
2.4.1.2	User-Centred Evaluation	29

2.4.2	Measures of Retrieval Effectiveness	32
2.4.2.1	Evaluating Ad-hoc Retrieval	33
2.4.2.2	Evaluating Diversity-Based Retrieval	36
II	Need for Diversity in Exploratory Search	40
3	Diversity-Based Recommendations for Image Browsing	41
3.1	Introduction	41
3.1.1	Goal and Plan of the Chapter	44
3.2	Background and Related Work	45
3.2.1	Exploratory Search	45
3.2.2	Image Retrieval and Browsing	46
3.2.3	Recommender Systems	50
3.3	System Description	52
3.3.1	System Overview	52
3.3.2	Interface Design	54
3.3.3	Recommendation Approach	56
3.4	Summary	59
4	User-Centred Evaluation of a Diversity-Based Recommender System	61
4.1	Introduction	61
4.2	Experiment and Validation	62
4.2.1	Research Questions	62
4.2.2	Experimental Assumptions and Scope of the Study	63
4.2.3	Plan of Experiments	64
4.2.3.1	Experimental Design	64
4.2.3.2	Collection and Data Pre-Processing	65
4.2.3.3	Search Tasks	67
4.2.3.4	Participants	68
4.3	Results and Analysis	69
4.3.1	User Perception	69
4.3.2	Usage Log File Analysis	73
4.4	Findings and Discussion	77

4.5	Summary	79
III	Ranking Paradigms for Result Diversification	81
5	Ranking Paradigms and their Integrations for Sub-topic Retrieval	82
5.1	Introduction	82
5.1.1	Inter-dependent Document Relevance Paradigm	83
5.1.2	Sub-topic Aware Paradigm	84
5.1.3	Goal and Plan of the Chapter	85
5.2	The Probability Ranking Principle	86
5.3	Background of Result Diversification	89
5.3.1	Beyond Independent Relevance	89
5.3.1.1	Maximal Marginal Relevance	90
5.3.1.2	Modern Portfolio Theory	91
5.3.2	Sub-topic Aware Paradigm for Diversity	92
5.3.2.1	Sub-topic Modelling Techniques	95
5.3.2.2	Post-Clustering Methods for Document Selection	97
5.4	Diversification with Two Ranking Paradigms	100
5.4.1	Motivation	100
5.4.2	Proposed Framework	101
5.4.3	Integration Approach	103
5.5	Summary	105
6	Empirical Study of Ranking Diversification for Sub-topic Retrieval	108
6.1	Introduction	108
6.2	Experiment and Validation	109
6.2.1	Research Questions	109
6.2.2	Experimental Assumptions and Scope of the Study	110
6.2.3	Plan of Experiments	111
6.2.3.1	Test Collections and Topics	111
6.2.3.2	Evaluation Measures	115
6.2.3.3	Experimental Systems and their Settings	116
6.3	Results and Analysis	119

6.3.1	Results in ImageCLEF 2009	120
6.3.2	Results in ClueWeb 2009	122
6.3.3	Results in TREC Interactive 6,7,8	124
6.4	Findings and Discussion	125
6.5	Summary	127
IV	Evaluation Measures in Sub-topic Retrieval	128
7	Diversity and Redundancy-Based Measures	129
7.1	Introduction	129
7.2	Evaluation Methodology in Diversity Task	131
7.3	User Models within the Diversity Task	132
7.3.1	User Model 1 – a set of users and a single document	133
7.3.2	User Model 2 – a single user and a set of documents	136
7.3.3	Recap of the Two User Models	138
7.4	Evaluation Measures	138
7.4.1	Diversity-Based Measures	139
7.4.1.1	Sub-topic Recall	139
7.4.1.2	Sub-topic Mean Reciprocal Rank	140
7.4.2	Redundancy-Based Measures	141
7.4.2.1	Novelty Biased Discounted Cumulative Gain	142
7.5	Analysis of α -nDCG	146
7.6	Deriving a Safe Threshold for α	150
7.7	Examination of the Safe Threshold	153
7.8	Summary	154
8	Evaluation of the Safe Threshold for α-nDCG	157
8.1	Introduction	157
8.2	Experiment and Validation	158
8.2.1	Research Questions	158
8.2.2	Experimental Assumptions	159
8.2.3	Plan of Experiments	159
8.3	Results and Analysis	163

8.3.1	A Real Case Example	163
8.3.2	Kinematics	165
8.3.3	Correlations	166
8.3.3.1	Real Systems	168
8.3.3.2	Synthetic Systems	169
8.3.4	Reliability of α -nDCG	172
8.3.4.1	Discriminative Power	172
8.3.4.2	Stability and Sensitivity on Swap Method	174
8.4	Findings and Discussion	177
8.5	Summary	178
9	Re-analysing Diversification Approaches for Sub-topic Retrieval	179
9.1	Introduction	179
9.2	Results and Analysis	180
9.2.1	Re-analysed Results of ImageCLEF 2009	180
9.2.2	Re-analysed Results of TREC ClueWeb 2009	182
9.2.3	Re-analysed Results of TREC 6,7,8 Interactive	184
9.3	Discussion and Conclusion	185
V	Conclusion	188
10	Conclusions	189
10.1	Summary of Work and Discussion	189
10.1.1	Diversity-Based Recommender System	189
10.1.2	An Integration Framework for Result Diversification	191
10.1.3	Query-Basis Approach to Derive Safe Threshold of α -nDCG	192
10.2	Contributions	194
10.3	Future Work	195
VI	References and Appendices	199
	References	221
A	Architecture and Implementation of <i>Ostensive Browser Plus</i>	222

B	Diverse Recommendations Through Image Browsing: Experimental Documents	225
B.1	Information Sheet	226
B.2	Consent Form	227
B.3	Task Descriptions	228
B.4	Entry Questionnaire	232
B.5	Post-Search Questionnaire for Baseline System	235
B.6	Post-Search Questionnaire for Recommender System	240
B.7	Exit Questionnaire	246
C	Images Features Implemented in OBP	248
C.1	Overview of Implemented Visual Features	248
C.2	Basic Image Data	248
C.3	Low-level Features	250
C.3.1	Colour Layout Descriptor	250
C.3.2	Edge Histogram Descriptor	252
C.3.3	Homogeneous Texture Descriptor	255
D	TREC 2010 Web Diversity Track Guidelines	257
D.1	TREC Guideline	258

List of Figures

3.1	Graph-based image browsing	49
3.2	Example of similarity-based image recommendations and diversity-based image recommendations	50
3.3	The components of the aspectual browsing system with recommendations	53
3.4	Browsing interface of the OBP system	54
3.5	Slide-show interface of the OBP system	56
4.1	Users' satisfaction after interacting with two systems (higher is better), as asked in post-search questionnaires.	71
4.2	Users' interaction patterns over the course of an experimental session when using two systems.	77
5.1	Re-ranking methods for promoting diversity.	93
5.2	Diversification with cluster ranking. The input is a ranked list of documents and output is a diversified ranked list of documents.	102
6.1	Example of ImageClef 2009 Photo Retrieval dataset entry.	112
6.2	Example of TREC ClueWeb 2009 dataset.	113
6.3	Example of TREC 6, 7, 8 interactive dataset.	114
7.1	Values of the safe threshold for α	152
8.1	The sub-topic distribution of TREC 2009 and 2010 queries and the relative percentage of queries, for which the setting $\alpha = st + 0.01$ produces different system rankings than the setting $\alpha = 0.5$	160

LIST OF FIGURES

8.2	Average performance of the systems participating at TREC 2009 and 2010 Web Diversity track, divided into nine performance categories. .	161
8.3	Kinematics of 48 system runs submitted to TREC 2009 on 39 queries, with respect to α -nDCG@10 when $\alpha=0.5$, and their movements against α -nDCG@10 with $\alpha > st$	166
8.4	Kinematics of 32 system runs submitted to TREC 2010 on 37 queries, with respect to α -nDCG@10 when $\alpha=0.5$, and their movements against α -nDCG@10 with $\alpha > st$	167
8.5	ASL curves based on Paired Bootstrap Hypothesis Tests on TREC 2009.	173
8.6	ASL curves based on Paired Bootstrap Hypothesis Tests on TREC 2010.	174
8.7	MR-PT curves of α -nDCG with the two different settings of α on TREC 2009 and 2010. This curves are used to assess the stability of the measure under the different settings.	175
C.1	RGB colour space.	249
C.2	Images representation using RGB colour model.	249
C.3	Five edge types to create EHD.	252
C.4	Definition of sub-image and image-blocks.	253
C.5	30 frequency channels used in computing the HTD.	255

List of Tables

2.1	Graeco-Latin square design of two systems and two tasks. Each row represents an order of system-task pairs, assigned to users to perform in an experiment.	31
4.1	The experimental design follows a Graeco-Latin square rotation for systems (<i>S1–S2</i>) and tasks (<i>T1–T4</i>), involving 24 users (<i>U1–U24</i>) . .	65
4.2	21 semantic differentials in post-search questionnaires	70
4.3	Users’ perception of comparing two systems in an exit questionnaire. .	73
4.4	User interaction statistics – mean, standard deviation (in bracket), and percentage increment of <i>S2</i> over <i>S1</i> (below). No statistical significance at 0.05 level has been found between two systems.	74
4.5	No. of images (in percentage), obtained from text search, browsing and in particular recommendations, exploited to define aspects and marked as relevant in a recommender system (<i>S2</i>).	76
6.1	Statistics of three experimental collections.	111
6.2	No. of topics and statistics of sub-topic in three collections.	112
6.3	Retrieval performances on the <i>ImageCLEF 2009 (Photo Retrieval)</i> collection with % of improvement over PRP. Parametric runs are tuned w.r.t. α -nDCG@10 ($\alpha = 0.5$). Statistical significances at 0.05 level against MMR, and MPT are indicated by * and † respectively.	121
6.4	Retrieval performances on the <i>TREC ClueWeb 2009</i> collection with % of improvement over PRP. Parametric runs are tuned w.r.t. α -nDCG@10 ($\alpha = 0.5$). Statistical significances at 0.05 level against MMR, and MPT are indicated by * and † respectively.	123

6.5	Retrieval performances on the <i>TREC 6,7,8 interactive</i> collection with % of improvement over PRP. Parametric runs are tuned w.r.t. α -nDCG@10 ($\alpha = 0.5$). No statistical significance is computed due to the limited number of topics.	124
7.1	A document/sub-topic matrix representing the relevance judgements of a query q made on four documents d_1, d_2, d_3, d_4 with respect to four sub-topics s_1, s_2, s_3, s_4 . A full dot (i.e. \bullet) in a cell (i, j) of the matrix represents the case where a document d_i has been found relevant to sub-topic s_j	133
7.2	Relevance judgements based on <i>User Model 1</i> . A document/user matrix representing the relevance judgements made on four documents (i.e. d_1, \dots, d_4) by ten users (i.e. u_1, \dots, u_{10}), issuing the same query q . Each user is interested in only one sub-topic, where their sub-topic relevance judgements correspond to the matrix presented in Table 7.1. A full dot (i.e. \bullet) in a cell (i, j) represents the case where document d_i has been found relevant by user u_j	134
7.3	Relevance judgements based on <i>User Model 2</i> . A document/user matrix representing the relevance judgements made on four documents (i.e. d_1, \dots, d_4) by users (i.e. u_{11}, u_{12}), issuing the same query q . Each user is interested in more than one sub-topics and their sub-topic relevance judgements correspond to the matrix presented in Table 7.1. A full dot (i.e. \bullet) in a cell (i, j) represents the case where document d_i has been found relevant to sub-topic s_j	136
7.4	Five documents relevant to the sub-topics of query 26, “lower heart rate”, from the TREC 2009 Web Diversity Track.	147
7.5	Corresponding evaluations of three imaginary system rankings for query 26 using α -nDCG, when $\alpha=0.5$ and an ideal ranking, of which dcng(r) are used for normalisation	148
7.6	Corresponding evaluations of three imaginary system rankings for query 26 using α -nDCG, when $\alpha=0.68$ and an ideal ranking, of which dcng(r) are used for normalisation	154

LIST OF TABLES

8.1	Top six submitted runs from TREC 2009 Web Diversity track on query 35, evaluated by α -nDCG@10 (<i>Middle</i>). System rankings with scores when $\alpha = 0.5$ (<i>Top</i>) and $\alpha > st$ (<i>Bottom</i>).	164
8.2	Kendall's τ and τ_{ap} between rankings of systems submitted to TREC 2009 and 2010 and evaluated with α -nDCG@10 and $\alpha=0.5$ or $\alpha > st$, or s-recall. <i>All</i> systems are considered.	168
8.3	Kendall's τ and τ_{ap} between rankings of systems submitted to TREC 2009 and 2010 and evaluated with α -nDCG@10 and $\alpha=0.5$ or $\alpha > st$, or s-recall. Only the top 10, 15 and 20 systems are considered.	169
8.4	Kendall's τ and τ_{ap} between rankings of systems evaluated with α -nDCG@10 and $\alpha=0.5$ or $\alpha > st$. <i>Synthetic</i> systems are considered.	170
8.5	The performance of five synthetic runs on query 21 wrt. α -nDCG with $\alpha = 0.5$ and $\alpha > st$, and maximum s-mrr	171
8.6	Pearson's correlation between the system rankings obtained by s-mrr and those obtained by the two settings of α -nDCG.	172
8.7	Discriminative power of traditional metrics at significant level=0.05.	174
8.8	Difference and sensitivity based on the swap method (swap rate $\leq 5\%$) using systems from the TREC 2009 and 2010 Web Diversity tracks.	176
9.1	Retrieval performances on the <i>ImageCLEF 2009 (Photo Retrieval)</i> collection with % of improvement over PRP. Parametric runs are tuned w.r.t. α -nDCG@10 ($\alpha > st$). Statistical significances at 0.05 level against MMR, and MPT are indicated by * and † respectively.	181
9.2	Retrieval performances on the <i>TREC ClueWeb 2009</i> collection with % of improvement over PRP. Parametric runs are tuned w.r.t. α -nDCG@10 ($\alpha > st$). Statistical significances at 0.05 level against MMR, and MPT are indicated by * and † respectively.	183
9.3	Retrieval performances on the <i>TREC 6,7,8 interactive</i> collection with % of improvement over PRP. Parametric runs are tuned w.r.t. α -nDCG@10 ($\alpha > st$).	184
C.1	Lists of low-level features implemented in OBP.	248
C.2	The representation of CLD.	251
C.3	Semantics of histogram bins of the EHD.	254

Abbreviations

• AP	Average Precision
• ASL	Achieved Significance Level
• AWT	(Java) Abstract Windows Toolkit
• CLD	Colour Layout Descriptor
• CLEF	Cross-Language Evaluation Forum
• CMY	Cyan, Magenta, Yellow Colour Model
• C-Precision	Combined-Precision
• DCT	Discrete Cosine Transform
• EHD	Edge Histogram Descriptor
• ERR	Expected Reciprocal Rank
• ERR-IA	Intent-Aware Expected Reciprocal Rank
• HCI	Human Computer Interaction
• HTD	Homogeneous Texture Descriptor
• IA	Intent-Aware Measures
• ImageCLEF	Image Retrieval in Cross-Language Evaluation Forum
• iPRP	Interactive Probability Ranking Principle
• IR	Information Retrieval
• LDA	Latent Dirichlet Allocation
• MAP	Mean Average Precision
• MAP-IA	Mean Average Precision
• MMR	Intent-Aware Maximal Marginal Relevance
• MPEG-7	Multimedia Content Description Interface
• MPT	Modern Portfolio Theory
• MR	Minority Rate
• MRR	Mean Reciprocal Rank
• nDCG	Normalized Discounted Cumulative Gain
• nDCG-IA	Intent-Aware Normalized Discounted Cumulative Gain
• α-nDCG	Novelty-Biased Normalized Discounted Cumulative Gain
• NIST	(U.S.) National Institute of Standards and Technology
• OB	Ostensive Browser

• OBP	Ostensive Browser Plus
• PLSA	Probabilistic Latent Semantic Analysis
• PRP	Probability Ranking Principle
• PT	Proportion of Ties
• qPRP	Quantum Probability Ranking Principle
• RGB	Red, Green, Blue Colour Model
• RR	Reciprocal Rank
• S-MRR	Sub-topic Mean Reciprocal Rank
• S-Recall	Sub-topic Recall
• ST	Safe-Threshold of α -nDCG
• TREC	Text Retrieval Conference
• WWW	World Wide Web
• XML	Extensible Markup Language
• YCbCr	Luma, Chroma-blue, Chroma-red Colour Model

Part I

Introduction and Background

Chapter 1

Introduction

The World Wide Web (or, Web) has matured as a ubiquitous platform for communication. Increasingly vast quantities of information are created and distributed, with studies suggesting that more than a billion new pages are added daily¹. The Web contains documents of diverse characteristics, including product descriptions, news and magazine media, academic papers, and encyclopaedic articles. Text, images, audio, and videos are just a few of the many media types present in these documents. The prevalence of “Web 2.0²” in the form of social and community-based websites such as personal blogs, Wikipedia, Facebook, and Twitter has meant many users have changed the way they engage with information. Rather than passively consuming information, many users are voluntarily creating their own (user-generated) content. Every 60 seconds, 20,000 new posts are published on the micro-blogging platform Tumblr and 98,000 tweets on Twitter³. As such, a great deal of information is now created by end users as well as traditional content providers. However, such information may contain highly similar or nearly duplicate content. This is because, for example, popular news is often repeatedly posted and discussed by different users or media outlets, leading to a great deal of overlapping content on the Web.

Since the advent of search engines⁴, information has become considerably quicker

¹<http://www.worldwidewebsize.com/>. From 26 websites in 1992, there are now at least a hundred million websites with 15.72 billion web pages estimated, updated: 1 June 2011.

²A popular term for advanced internet technology and web applications, which facilitate users to collaborate and share information online.

³<http://www.go-gulf.com/blog/60-seconds>

⁴The practical application of the principles of information retrieval, often applied to large-scale document collections.

and easier for individuals to access. Modern search engines have provided sophisticated tools to resolve users' *information needs*, locating relevant information based on the user's description of their request (e.g. query). Search engines have hitherto played a vital role in seeking useful information in a world of unprecedented information overload. Nevertheless, on the Web there may exist a high number of documents being relevant to users but containing very similar content. Hence, search engines that treat documents independently are likely to retrieve documents with the same or similar content. As such, subsequent redundant (despite relevant) documents in the search results may be considered less useful or at most non-relevant if users have already examined other documents containing the same information. A challenging problem for search engines is to retrieve not only *relevant* but also *diverse* documents.

In information-seeking activities, one of the tasks users usually engage in is exploratory search. Without any prior domain knowledge of the search topics, users may have no clear steps towards finding relevant information in the information space. They are likely to be uncertain about which query they should submit at the beginning of their search, what types of documents are present in the collections, and thus how to reach the information they are seeking [Salton and McGill, 1986]. Users engaged in exploratory search usually pose *broad* or *tentative* queries so as to navigate through the information space proximal to the relevant information. They then perform a combination of searching and browsing activities to explore information and learn how to exploit it. Once users perceive and internalise information, their newly acquired knowledge is used to address problems regarding the search topics. It can be argued that retrieval is necessary but not sufficient when information is sought to address human curiosities related to information exploration, such as in scientific discovery, learning, decision-making, etc. [White and Roth, 2009]. Therefore, users do require additional supports for learning search domains so as to clarify their goals and actions [Newell and Simon, 1972; Simon, 1973]. To leverage exploratory search, retrieval systems have to find a way of hedging bets on choosing what to return; that is a broad view of possible search *aspects* that can support information seeking requests.

Answering ambiguous or underspecified queries is another main challenge for search engines in retrieving relevant documents [Clarke et al., 2009b]. Consider the following example scenario: a user issues the query “apple” to a Web search engine and scans the retrieved search results. The first result refers to the Apple Inc. home

page. It might contain the sought information, but the user is not certain and moves on. The second result links to the home page of the Apple online store. Since the previous result was considered not relevant, this page is equally unlikely to be relevant. The third result is the Wikipedia page about apple as a fruit. No, definitely not relevant to the user at all. The user eventually lands to the forth page, a technology article, providing most of the sought information: rumours about an upcoming Apple’s operating system, which has not yet officially announced and advertised by Apple Inc. The user clicks on this link and never returns to the result list again. Of course, as illustrated by the example, although the first query is ambiguous and underspecified, the user still gets benefits from a search system that diversifies results about the company, the fruit, the operating system and other senses of the query “apple”. When generating a ranked result, a search system should attempt to maximise the probability that a user will obtain the sought information.

In order to facilitate information seeking activities, there has been growing interest in building and optimising information retrieval (IR) systems that provide relevant and diverse information in a unified manner: this area of research is called “**diversity-based retrieval**” or “**sub-topic retrieval**” [Carbonell and Goldstein, 1998; Zhai et al., 2003]. Relevance and diversity together can potentially increase the usefulness of IR systems as perceived by users. In generating a result list, the IR systems must supply novel relevant information as users traverse the list themselves, cover all possible aspects of the needs underlying a query or an interaction with the systems, and balance the diverse needs of the entire user population.

1.1 Diversity in Information Retrieval

The focus of this thesis is on the topic of diversity in information retrieval. A question that generally comes up with the need to cater for diversity is: *what is diversity?* Clarke et al. [2008] and Zhai et al. [2003] addressed the need for diversity in search results and categorised it into two groups: either an inherent property of information need(s) (to promote novelty) or a dynamic property determined by the users of a system (to address uncertainty of information need). In this thesis, we aim to overcome the problems associated with these two groups of diversity, i.e. eliciting the need of diversity in exploratory searchers and promoting diversity to avoid redundancy. Similarly in

the *SIGIR 2009 Workshop on Redundancy, Diversity, and Inter-dependent Document Relevance* [Radlinski et al., 2009], the precise distinctions between the two categories were defined as follows:

- 1) Intrinsic diversity
- 2) Extrinsic diversity

1.1.1 Intrinsic Diversity

Diversity is considered a property of information need since there is no single result that can satisfy user information need. This type of diversity aims to find a *set* of different results, which together fulfil a *single* well-defined information need. Diversity can be seen as a means to avoid redundancy in search results because presenting documents that contain similar information may not benefit users. An example in this category is an information need, which requires a set of two or more answers. A user would wish for diverse results to obtain the overview of a search topic and increase the confidence about the clearness and correctness of the answer for an information need. Another user would desire different aspects of a topic, such as a variety of reviews about a product, or a variety of opinions about a political issue. Clarke et al. [2008] consider this type of diversity as optimising for *novelty* in search results, where the goal is to retrieve all different aspects in order to fulfil the user information need.

1.1.2 Extrinsic Diversity

Diversity is considered as a means to deal with the uncertainty about a user's information need. This type of diversity can be further divided into three subgroups. First, the uncertainty can come from the poor representation of information needs, commonly expressed by queries. Although user's information needs are clear or well-defined, the issued query might be linguistically ambiguous and can be interpreted into two or more distinct meanings (i.e. polysemy). This is not only due to the multiple interpretations of the query, but also due to the ambiguity in the named entities or acronyms the query refers to. For instance, the query "house" may mean "home", "building", or "assembly", whereas the request "victoria" may refer to "person", "place", or "brand".

Second, the uncertainty can derive from different user interests in the information space, where users who have different backgrounds may refer to different aspects of a search topic. A simple example is the unambiguous query “house plans” that may refer to different aspects depending on the views of users (i.e. “technical drawings” for architects or “creative designs” for customers). These first two subgroups of extrinsic diversity are intimately related to the query formulation problem. Users have difficulty to formulate a good query that results in the retrieval of relevant documents. This is normally because the user’s query is *ambiguous* or *underspecified*. Alternatively, a retrieval system may suggest a new query for better specifying information need such as word sense disambiguation or query expansion. Instead, result diversification aims to alleviate this problem by providing a single “entry-point” result page which contains all possible senses of results that users are seeking.

Third, the uncertainty can stem from the inherent incompleteness [Ingwersen, 1992] of a user’s information need. In this case, searches are motivated by information needs that are genuinely unclear and uncertain. In other words, users are either unsure about their goal at the beginning of their search or unfamiliar with the domain of their goal (i.e. need to learn about the topic so as to understand how to achieve the goal). This type of search is referred to as the exploratory search [White and Roth, 2009], which includes information-seeking activities, for example, the acquisition of knowledge or the development of intellectual skills. Users perform a bundle of search activities, e.g. querying, browsing, clicking, etc., so as to develop their cognitive capabilities and clarify their information needs. Thus, a requirement for the diversity is to allow users to explore, learn, and opt for information that can ultimately be used to clarify their uncertain information needs.

1.2 Thesis Statement

Overall, this doctorate work is an exploration into three territories of diversity-based document retrieval; recommender systems, retrieval algorithms, and evaluation measures. First of all, this thesis aims to understand and describe the need for diversity in search results from the users’ perspective. To this end, we conduct a user study that considers an interactive recommender system, called Ostensive Browser Plus (OBP) [Urban et al., 2006], which is designed for exploratory search tasks. It is

assumed that users, who are engaged in exploratory search, may prefer systems that provide diverse results over traditional search systems. We introduce the OBP system featured with diverse recommendations, aspectual interfaces, and content-based browsing. By incorporating implicit relevance feedback and clustering techniques, the recommender system can adaptively provide relevant and diverse documents to satisfy evolved user information need. The aspectual interface is also introduced in the system so that users can structure their search, helping them to discover more aspects of a search topic. We claim that this recommender system provides diverse recommendations that can support users in exploratory work tasks.

Secondly, we review and study the state-of-the-art approaches for automatic result diversification. These approaches can be classified into two categories, i.e. inter-dependent document relevance and the sub-topic aware paradigms. The former focuses on “implicit” diversity or novelty defined in terms of previously ranked documents whereas the latter focuses on “explicit” diversity by directly using (sub-)topical categories predicted from documents or query logs. We develop a ranking framework based on the integration of the two paradigms and empirically investigate the best technique for combining them. We prove that the integration approach has potential to improve the coverage of sub-topics given a query, especially when sub-topics are predicted with a high quality.

Finally, this thesis studies evaluation measures in the context of diversity-based retrieval. We identify an issue with the de-facto standard measure, novelty-biased discounted cumulative gain (α -nDCG) [Clarke et al., 2008]. The issue causes the measure to misbehave, i.e. favour retrieval systems that present *redundant* information rather than systems that provide *novel* and *diversified* information. Recognising this characteristic of the measure is of importance since it affects the evaluation of retrieval systems. To overcome this problem, we derive a theoretically sound solution by defining a safe threshold for the measure on a query-basis. We prove that the diversity of document rankings can be intuitively measured by employing our proposed safe threshold.

1.2.1 Research Questions

The following research questions were investigated throughout this thesis:

- **RQ1:** How effectively does the diversity in search results support users who are engaged in exploratory search? Do the users perceive that diversified results are useful when they have information needs related to multiple aspects?
- **RQ2:** How can implicit relevance feedback be used for generating *relevant* and *diverse* recommendations?
- **RQ3:** What are the differences between ranking approaches for result diversification? Can these approaches be categorised in order to provide reasoning about their ranking patterns in general?
- **RQ4:** How can we model a new ranking framework that improves the effectiveness of existing diversification approaches?
- **RQ5:** Do current evaluation measures in diversity retrieval actually assess the utility of document rankings in terms of relevance, novelty, and in particular diversity? If not, can we devise an approach to resolve this issue?
- **RQ6:** Does our proposed approach to resolve the issue change the evaluation measure to assess document rankings according to the goal of the diversity retrieval task? How reliable is our approach to evaluate the performance of diversity-based retrieval systems?

1.3 Contributions

The main contributions of this work can be summarised as follows:

- **C1:** An investigation of the need for result diversity from the users' point of view is introduced. A user-centred experiment is conducted under exploratory search conditions: e.g. search modality, systems, and tasks.
- **C2:** A diversity-based recommender system is introduced. The system provides diverse recommendations mined from user's implicit relevance feedback. The feedback is derived from the user's interactions with a system while browsing a collection. The aspectual browsing interface is also firstly introduced to the content-based browsing system.

- **C3:** Ranking paradigms in result diversification are classified, and this is crucial for the development of new ranking strategies. At the time of writing, it was the first contribution that empirically studied different ranking paradigms in a number of experimental contexts. Therefore, this work provides insights into method development and future directions for the research in this area.
- **C4:** An effective general framework that integrates the inter-dependent document relevance and the sub-topic aware paradigm is devised. The framework is proposed to improve the effectiveness of ranking diversification so as to cover more aspects of a query.
- **C5:** A comprehensive analytical and empirical investigation of the *de-facto* standard measures, in particular α -nDCG, in sub-topic retrieval. We disclose a situation, where arbitrarily setting a parameter α of α -nDCG causes the measure to deviate from the desired outcome of the evaluation context.
- **C6:** A theoretically sound solution that determines a safe threshold for α value on a query-basis is introduced to effectively measure performance of systems that promote diversity.
- **C7:** In a wider perspective, we show that the notion of diversity spans the areas of exploratory search, document redundancy, query ambiguity, and uncertainty about users.

1.4 Roadmap of the Thesis

This thesis is structured into five main parts, which contain the corresponding chapters:

- **Part I: Introduction and Background**

This part comprises of two chapters. It introduces the concept of diversity-based retrieval and provides the background material for this thesis. The outline and overall aim of the thesis is provided in Chapter 1. Next, related research and fundamental concepts in IR are revisited in Chapter 2. We also describe basic problems of diversity tasks as well as evaluation methodologies and performance measurements widely used in IR context.

- **Part II: Diversity-Based Recommendations for Image Browsing**

In this part, we investigate the need for result diversity from the users' perspective. We propose a diversity-based image recommender system for the purpose of the investigation. The recommendation is mined from implicit relevance feedback extracted from user browsing trails and diversified based on image content. An aspectual browsing interface, which allows a user to define and organise their own search aspects is also implemented in the system. The system as a whole is developed to facilitate exploratory search tasks. Chapter 3 illustrates the system and its interface as well as its graph-based algorithm for aggregating user's implicit relevance feedback for recommendation. User-centred evaluation is performed and discussed in Chapter 4.

- **Part III: Ranking Paradigms for Result Diversification**

This part begins by presenting an overview of re-ranking approaches and strategies for promoting diversity in a result list. These approaches are sub-divided into two main paradigms according to their ranking strategies: *i*) inter-dependent document relevance and *ii*) diversification based upon topic modelling. In Chapter 5, we introduced a *new* diversification framework, which integrates the two paradigms to enhance the performance of result diversification. An empirical study on comparing and combining the two ranking paradigms in the context of diversity retrieval is conducted and reported in Chapter 6.

- **Part IV: Evaluation Measures in Sub-topic Retrieval**

In this part, we revisit evaluation measures in diversity-based document retrieval. We highlight the problem of setting a parameter α in α -nDCG. We emphasise that setting an arbitrary value of α prevents the measure from behaving as desired, i.e. assessing the effectiveness of systems that provide complete coverage of the query-intents by avoiding excessive redundancy. Chapter 7 unveils our approach to overcome this situation by defining a safe threshold for α . A comprehensive evaluation on both TREC submissions and synthetic data is performed and presented in Chapter 8. Finally, we re-analyse all the diversification approaches conducted in Chapter 6 using α -nDCG with our proposed parameter setting. The re-analysed results are reported in Chapter 9.

- **Part V: Conclusion**

The summary of this thesis is discussed in Chapter 10, involving three aspects of our studies in diversity topics (i.e. system, modelling, and evaluation). We highlight the contributions of our works from the results of experiments shown in Part II, III, and IV.

The thesis comprises four appendices. The first (Appendix A) describes the architecture and implementation of the diversity-based recommender system. Appendix B provides all questionnaires and information sheets handed out in a user study (Chapter 3). Appendix C presents a brief implementation technique for visual feature extraction and similarity matching, employed in the recommender system. Note that the study of visual features is not one of the main objectives of this thesis. Finally, Appendix D reports a guideline of TREC 2010 Web Diversity Track that stresses the requirements of systems and effectiveness measures in diversity-based retrieval.

1.5 Publications

The research and results presented in this doctoral thesis are mainly included in the following publications:

- [Leelanupab et al., 2011]: **A Query-Basis Approach to Parametrizing Novelty-Biased Cumulative Gain**, T. Leelanupab, G. Zuccon, and J.M. Jose, the 3rd International Conference on Theory of Information Retrieval: Advances in Information Retrieval Theory, ICTIR 2009
- [Leelanupab et al., 2010d]: **When Two is Better than One: A Study of Ranking Paradigms and Their Integrations for Sub-topic Retrieval**, T. Leelanupab, G. Zuccon, and J.M. Jose, the 6th Asia Information Retrieval Societies Conference, AIRS 2010
- [Leelanupab et al., 2010c]: **Technical Report: A Study of Ranking Paradigms and Their Integrations for Sub-topic Retrieval**, T. Leelanupab, G. Zuccon, and J.M. Jose, Technical Report, School of Computing Science, University of Glasgow, 2010

- [[Leelanupab et al., 2010b](#)]: **Revisiting Sub-topic Retrieval in the ImageCLEF 2009 Photo Retrieval Task**, T. Leelanupab, G. Zuccon, and J.M. Jose, Chapter 15 in ImageCLEF – experimental evaluation in image retrieval, Springer Berlin Heidelberg, 2010
- [[Leelanupab et al., 2010a](#)]: **University of Glasgow at ImageCLEFPhoto 2009: Optimising Similarity and Diversity in Image Retrieval**, T. Leelanupab, G. Zuccon, A. Goyal, M. Halvey, P. Punitha, and J.M. Jose, Multilingual Information Access Evaluation II. Multimedia Experiments, Springer Berlin Heidelberg, CLEF 2009
- [[Leelanupab et al., 2009a](#)]: **A Simulated Evaluation of Image Browsing Using High-Level Classification**, T. Leelanupab, Y. Feng, V. Stathopoulos, and J.M. Jose, the 4th International Conference on Semantic and Digital Media Technologies: Semantic Multimedia, SAMT 2009
- [[Leelanupab et al., 2009b](#)]: **Application and Evaluation of Multi-Dimensional Diversity**, T. Leelanupab, M. Halvey, and J.M. Jose, the Theseus/ImageCLEF workshop on visual information retrieval evaluation, 2009
- [[Zuccon et al., 2009b](#)]: **The University of Glasgow at ImageClefPhoto 2009**, G. Zuccon, T. Leelanupab, A. Goyal, M. Halvey, P. Punitha, and J.M. Jose, Image Retrieval in CLEF 2009, ImageCLEF 2009
- [[Leelanupab et al., 2009c](#)]: **User Centred Evaluation of A Recommendation Based Image Browsing System**, T. Leelanupab, F. Hopfgartner, and J.M. Jose, the 4th Indian International Conference on Artificial Intelligence, IICAI 2009

other related publications are:

- [[Whiting et al., 2011b](#)]: **University of Glasgow (qirdcsuog) at TREC Crowdsourcing 2011: TurkRank – Network-based Worker Ranking in Crowdsourcing**, S. Whiting, J. Rodriguez Perez, G. Zuccon, T. Leelanupab, and J.M. Jose; Text REtrieval Conference, NIST, 2011, to appear

- [[Zucon et al., 2011c](#)]: **Crowdsourcing Interactions: Capturing Query Sessions through Crowdsourcing**, G. Zucon, T. Leelanupab, S. Whiting, E. Yilmaz, J.M. Jose, and L. Azzopardi, ECIR Workshop on Information Retrieval over Query Sessions, SIR 2011
- [[Zucon et al., 2011b](#)]: **Crowdsourcing Interactions: A Proposal for Capturing User Interactions through Crowdsourcing**, G. Zucon, T. Leelanupab, S. Whiting, J.M. Jose, and L. Azzopardi, WSDM Workshop on Crowdsourcing for Search and Data Mining, CSDM 2011
- [[Elliott et al., 2009](#)]: **An Architecture for Life-long User Modelling**, D. Elliott, F. Hopfgartner, T. Leelanupab, Y. Moshfeghi, and J.M. Jose, UMAP Workshop on Life-long User Modelling, LLUM 2009
- [[Leelanupab and Jose, 2008](#)]: **An Adaptive Browsing-Based Approach for Creating a Photographic Story**, T. Leelanupab, and J.M. Jose, the 3th International Conference on Semantic and Digital Media Technologies, SAMT 2008

Chapter 2

Retrieval Models, Tasks, and Evaluation

2.1 Introduction

In this chapter, we provide background information and definitions that are essential to understand the rest of this thesis. Instead of presenting an exhaustive survey of all related work here, we provide a complete account of related work applicable to the work of each part in the subsequent chapters, where it is easier to put into context and compare against our work. Topics covered in this chapter comprise basic information retrieval concepts, an overview of diversity-based document retrieval, and evaluation frameworks.

Fundamental concepts of information retrieval are introduced in Section 2.2. Section 2.3 outlines information retrieval tasks, focusing on diversity-based retrieval. In Section 2.4, we discuss various evaluation methodologies and effectiveness measures that have been established in the information retrieval domain.

2.2 Fundamental Concepts of Information Retrieval

In the context of information retrieval (IR), one of the primary goals is to understand and formalise the processes by which humans assess the relevance of documents with respect to their information needs. To understand human decisions on relevance it would probably be necessary to understand how languages are represented and processed in human brains; however, we are a long way from formalising this [Manning

et al., 2008]. Instead, IR researchers proposed theories about relevance, usually in the form of mathematical models, to address and specify how documents and information needs (i.e. queries) are represented and matched. These models provide algorithms and criteria, known as ranking functions, to estimate the relevance of documents with regard to queries. In other words, ranking functions allow quantification of the similarities amongst documents and queries. The estimates of document relevance are then employed to present *relevant* documents near the top of the ranking. Luhn [1957, 1958] provided the foundation for IR models, in which the unified representation of documents and queries, as well as the application of term weighting, is given. His work is often deemed as the precursor to *tf.idf* and related weighting schemes.

2.2.1 Boolean Model

A very simple model for document retrieval is the Boolean model for IR [Baeza-Yates and Ribeiro-Neto, 1999; Salton et al., 1983]. As it is conceptually based on set theory, the Boolean model indexes documents by considering the absence or presence of keywords or terms in the documents. As a result, the weights of index terms are represented in a binary format, TRUE or FALSE (i.e. 1 or 0). Likewise, binary weights are assigned to the index terms of queries, which can be formed and linked together by Boolean logical operators (e.g. AND, OR, NOT). Thus, a query is a conventional Boolean expression, resulting in only two possible outcomes for query evaluation (i.e. relevant or not-relevant). The Boolean retrieval model is hence known as *exact-match retrieval* since documents are retrieved only if they exactly match the query specification, otherwise they are not retrieved. Although this defines a clean formalism and simplicity behind ranking and term weighting, Boolean retrieval is not generally described as a ranking algorithm. This is because it assumes that all documents in the retrieved set are equivalent in terms of relevance. As a result, the retrieved documents will be presented to users in some order (e.g. creation date, author, or authority) regardless of their actual relevance with respect to a query. This is a major drawback of this approach, preventing good retrieval performance since it makes no distinction between the first document in a result list and the other retrieved documents.

Besides, there is no partial matching to the query specification. For example, let W be a set of vocabulary words in an entire collection consisting of w_1, w_2, w_3 , and d_i

be a document containing only a single index term w_1 . Thus, a document d_i can be represented by $d_i = (1, 0, 0)$. Assuming a query $q = (w_1 \vee w_2) \wedge w_3$, the document d_i is considered not-relevant in the Boolean retrieval model. Although the document d_i includes the term w_1 as specified by the query q , it will not be retrieved since it partially matches the query condition. Due to the need of exact matching and the lack of a sophisticated ranking algorithm, Boolean retrieval is no longer considered as a state-of-the-art ranking method [Zobel and Moffat, 2006].

Despite the above negative aspects, there are some benefits to the Boolean model. For instance, from an implementation point of view, Boolean retrieval is usually more efficient than ranked retrieval. This is because documents are rapidly eliminated from consideration in the scoring process if they do not match the query specification.

2.2.2 Vector Space Model

The vector space model was the basis for most of the research in IR in the 1970s. It was proposed to avoid the limitations of the Boolean model, in particular to make partial matching possible [Salton and McGill, 1986; Salton et al., 1975]. In the vector space model, documents and queries are viewed as points on a t -dimensional space of Euclidean geometry, where t is the number of index terms (words, stems, phrases, etc.). In general, the dimension of Euclidean vector space that spans the entire document collection is equal to the number of index terms contained in that collection. These t terms represent all the document features that are indexed by a retrieval system. A document d_i is represented by a vector of index terms, i.e. $\mathbf{d}_i = (d_{i,1}, d_{i,2}, \dots, d_{i,t})$, where $d_{i,j}$ is the weight of the j -th term in the document. Similarly, a query q is represented by $\mathbf{q} = (q_1, q_2, \dots, q_3)$, where q_j is the weight of the j -th term in the query. In the vector representation of documents and queries, non-binary weights are assigned to index terms. These term weights are eventually employed to compute the degree of similarity between each document in a collection and a query posed by a user. The retrieved documents are then able to be ranked according to the degree of similarity, such as in decreasing order of similarity score.

To compute the similarity between each document and a query, the Euclidean distance between points that represent each document and a query may be used. Nevertheless, a similarity measure (instead of a distance or dissimilarity measure) is more

2.2 Fundamental Concepts of Information Retrieval

commonly employed, so that the highest scored documents are the most similar to a query. A number of similarity measures have been proposed in literature and a survey of these measures is provided by [van Rijsbergen \[1979\]](#). The most popular of them is the *cosine correlation* similarity measure, known as the *cosine similarity*. The cosine similarity assesses the cosine of the angle θ between two vectors. Thus, the cosine correlation between \mathbf{d}_i and \mathbf{q} is defined as:

$$\text{sim}(d_i, q) = \cosine \theta_{\mathbf{d}_i, \mathbf{q}} = \frac{\mathbf{d}_i \bullet \mathbf{q}}{\|\mathbf{d}_i\| \|\mathbf{q}\|} = \frac{\sum_{j=1}^t d_{i,j} \times q_j}{\sqrt{\sum_{j=1}^t d_{i,j}^2 \times \sum_{j=1}^t q_j^2}}$$

The numerator of the cosine measure is the sum of the products of the term weights in a document and query (known as the dot product or inner product). The denominator normalises the score of dot product by dividing by the product of the lengths of the two vectors. There is no specific reason why the cosine correlation is preferred to other similarity measures, but it performs somewhat better in evaluations of search quality [[Croft et al., 2009](#)].

The values of the elements of each vector depend on the weighting scheme that is employed. Index term *weights* reflect importance of respective terms in a document and collection. Many different weighting schemes have been proposed based on retrieval models. One of the most common weighting schemes is *tf.idf* weighting. There are many variations of this weighting, but they are all rooted in a combination of the count of index term occurrence in a document, called *term frequency* (*tf*) and the frequency of index term occurrence over the entire document collection, called *inverse document frequency* (*idf*). The *tf* component reflects the importance of a term in a document whereas the *idf* reflects the importance of the term in a document collection. The *tf.idf* weight of term k in a document d_i is usually computed as follows:

$$tf_{i,k} \cdot idf_k = \frac{f_{i,k}}{\sum_{j=1}^t f_{i,j}} \cdot \log \frac{N}{n_k}$$

where $f_{i,k}$ represents the number of occurrences of the term k in the document d_i , N is the number of documents in the collection, and n_k is the number of documents in which term k appears at least once.

2.2.3 Probabilistic Models

Probabilistic retrieval models are currently the dominant ranking paradigm in information retrieval. They all are rooted in the *Probability Ranking Principle* (PRP) [Robertson, 1977; Robertson and Spärck-Jones, 1976], which is well established on the foundation of probability theory. PRP was proposed to represent and manipulate the uncertainty that is inherent in the information retrieval process. Users start with *information needs* translated into *query representations*. Similarly, there are documents converted into *document representations*. Based on these two representations, an IR system attempts to determine how well documents satisfy the users' information needs. Two types of uncertainty arise when estimating documents for retrieval. Considering only a query, the IR system has an uncertain understanding of the information needs. Considering both query and document representations, the system has the uncertainty of estimating whether a document contains information relevant to the information need. Therefore, PRP provides a principled foundation for reasoning this uncertainty, which is addressed by estimating the probability of how likely a document is relevant to the information need. This is the intuition underlying the probabilistic models of information retrieval and is why the relevance of a document to a query is assessed probabilistically. The complete discussion of PRP is given in Chapter 5, where we argue that the PRP's assumption may lead to inappropriate rankings for diversity-based document retrieval.

In probabilistic retrieval models, document retrieval is viewed as a *classification* problem where the goal is to decide whether a document belongs to the relevant set or the non-relevant set. That is, IR systems based on the probabilistic models should classify the document as relevant or non-relevant, and retrieve it if it is relevant. Assuming that $\mathcal{R} \in \{R, \bar{R}\}$ is a binary variable with value either R corresponding to the relevant set, or \bar{R} corresponding to the non-relevant set, we are interested in computing the probability of relevance given a document d_i , i.e. $P(\mathcal{R} = R|d_i)$, or in short $P(R|d_i)$. Similarly, the *conditional* probability representing the probability of non-relevance is $P(\bar{R}|d_i)$. It would be reasonable to classify the document to the set in which it obtains the higher probability. In other words, a document d_i is considered relevant if $P(R|d_i) > P(\bar{R}|d_i)$ or $P(d_i|R)P(R) > P(d_i|\bar{R})P(\bar{R})$ when derived following the *Bayes' Rule*¹. This is the

¹ $P(R|d_i) = P(d_i|R)P(R)/P(d_i)$

2.2 Fundamental Concepts of Information Retrieval

same as classifying a document as relevant if:

$$\frac{P(d_i|R)}{P(d_i|\bar{R})} > \frac{P(\bar{R})}{P(R)} \quad (2.1)$$

The left-hand side of the equation (2.1) is known as the *likelihood ratio*, which retrieval systems employ as a score for ranking documents. That is, the highly ranked documents will be those that have a high likelihood of belonging to the relevant set. The following derivation is valid, where \propto indicates rank equivalence:

$$P(R|d_i) \propto \frac{P(d_i|R)}{P(d_i|\bar{R})} \quad (2.2)$$

Assume that the terms present in a document d_i , i.e. $\{d_{i,1}, d_{i,2}, \dots, d_{i,t}\}$, are conditionally independent, where $d_{i,j}$ is the weight of the j -th term in the document and t is the total number of terms present in the document. Then the equation (2.1) can be rewritten as:

$$\frac{P(d_i|R)}{P(d_i|\bar{R})} \approx \prod_{j=1}^t \frac{P(d_{i,j}|R)}{P(d_{i,j}|\bar{R})} \propto \sum_{j=1}^t \log \frac{P(d_{i,j}|R)}{P(d_{i,j}|\bar{R})} \quad (2.3)$$

Therefore, probabilities assigned to documents indicate their likelihood of being relevant to a user's information need. These are indeed determined by the probability of drawing terms, which compose each document, from the relevant and non-relevant classes. Several approaches have been proposed to estimate this probability, such as the 2-Poisson model [Bookstein and Swanson, 1974], the Binary Independence model [Robertson and Spärck-Jones, 1976], and the Okapi BM25¹ model [Robertson et al., 1994]. Amongst other models, we highlight the well-known BM25 model that was introduced by Robertson and Spärck-Jones [1976].

¹Okapi refers to the name of the information retrieval system that first implemented the BM25 weighting function at London's City University. BM stands for Best Match and 25 is a numbering scheme to keep track of weighting variants, as experimented by Robertson and Walker [1994].

2.2.3.1 BM25 Model

Based on the probabilistic retrieval model, BM25 was developed to include document and query term weights. In this approach, the weight w of term j in a document d_i is computed as:

$$w = \log \frac{(r_j + 0.5)/(R_q - r_j + 0.5)}{(n_j - r_j + 0.5)/(N - n_j - R_q + r_j + 0.5)} \quad (2.4)$$

where:

- r_j is the number of *relevant* documents containing term j ;
- n_j is the number of documents containing term j ;
- R_q is the number of *relevant* documents for a query q ;
- N is the number of total documents in the collection.

Note that r_j and R_q are set to zero if there is no relevance information (i.e. when r_j and R_q are unknown or cannot be determined a priori). In BM25, $P(R|d_i)$ is approximated as:

$$P(R|d_i) \propto \sum_{j \in q} w \cdot \frac{(k_1 + 1) \cdot f_{i,j}}{K + f_{i,j}} \cdot \frac{(k_2 + 1) \cdot qf_j}{k_2 + qf_j} \quad (2.5)$$

where w ¹ is given by the equation (2.4); $f_{i,j}$ is the frequency of term j in the document d_i ; and qf_j is the frequency of the term j in the query; and k_1 , k_2 , and K are parameters whose values are set empirically. For example, in TREC experiments the typical value for k_1 is 1.2 and for k_2 is in the range of 0 to 1,000 [Croft et al., 2009]. The constant k_1 determines how the *tf* component of the term weight changes as $f_{i,j}$ increases whereas the constant k_2 has a similar role in the query term weight. The parameter K normalises the *tf* component by the document length. In particular,

$$K = k_1 \cdot ((1 - b) + b \cdot \frac{dl_i}{avdl}) \quad (2.6)$$

¹The *idf* component, where the probabilistic nature of BM25 appears.

where b is a parameter, dl_i is the length of the document d_i , and $avdl$ is the average length of documents in the collection. The constant b regulates the impact of the length normalisation and is typically set to 0.75 in TREC experiments.

2.3 Information Retrieval Tasks

The purpose of IR systems is to retrieve relevant documents to satisfy users' information needs. Nevertheless, in many retrieval scenarios the information needs may be uncertain or vary depending on the goals and intentions of the users. For example, whereas the information need of a lawyer is to search for all pertinent case files, that of a typical web surfer is to look for general information on a topic. The information-seeking behaviour of an *exploratory* searcher is to find many different aspects of a search topic. In addition, queries that represent information needs may be ambiguous and/or underspecified. Therefore, the differences in these retrieval scenarios determine different IR tasks as well as evaluation methodologies and measures.

In this section we shall outline two information retrieval tasks: ad-hoc document retrieval and diversity-based document retrieval, where the latter is the focus of this thesis. Many other tasks are examined in information retrieval domain; examples include enterprise search (e.g. [Hawking, 2004]), patent search (e.g. Fujii et al. [2004]), information filtering (e.g. [Robertson and Soboroff, 2002]), information distillation (e.g. [Yang et al., 2007]), biomedical (genomic) search (e.g. [Roberts et al., 2009]), etc., where each task is characterised by its own evaluation framework and measures.

2.3.1 Ad-hoc Document Retrieval

Ad-hoc document retrieval is the most commonly studied task in IR, and recently is investigated in TREC 2009-11 Web tracks [Clarke et al., 2009a, 2010]. The goal of this task is to return documents that are relevant to an immediate and single-aspect information need, which is expressed in the form of a query. The returned documents are presented in a list and should be ranked in the order that retrieval systems believe they are most likely to match a given query. The relevance of a document is considered *independent* to that of other documents, which appear before it in the result list. In the ad-hoc retrieval, a user is assumed to be interested in all documents that satisfy

such an information need [Voorhees and Harman, 2005]. The user model of this task prescribes that a user examines retrieved documents in sequential order from the top of the ranking to a cut-off position r . Documents in rank positions greater than r are considered not retrieved. The goal of the task and its user model are reflected in the measures, which are used to evaluate retrieval systems. The evaluation measures for ad-hoc document retrieval will be outlined in Section 2.4.2.1.

2.3.2 Diversity-Based Document Retrieval

The task of diversity-based document retrieval stems from the need for retrieval systems to provide complete coverage of relevant *aspects* or *sub-topics* (also called intents or query-intents¹), each related to a different information need. The terms, aspects and sub-topics, are closely related. “Aspect” is a common term used to describe the problem of aspect retrieval in the TREC interactive track [Over, 2001]. This track defines an aspect as one of many possible answers to a question of a search task, where interactive searchers have to find and explore all the answers to satisfy their information need. In other words, in this task, user satisfaction is not only achieved by retrieving relevant documents, but these documents also have to contain different answers to the same question. Many researches in exploratory search also exploits the interactive test collection for the evaluation of interactive IR systems [White et al., 2008].

The definition of “sub-topic” is more typically used in sub-topic retrieval [Zhai et al., 2003], which has been studied in many IR evaluation workshops (e.g. TREC 2009–11 Web Diversity tracks, ImageCLEF 2008–09 (diversity) photo retrieval tasks). The problem of sub-topic retrieval is complex and context dependent. It mainly deals with the ambiguity of queries and the user’s uncertainty about a query that can refer to many aspects [Clarke et al., 2009a, 2010]. These aspects are associated with the user intents behind a query. In fact, user queries often carry some degree of ambiguity per se [Spärck-Jones et al., 2007]. Genuinely ambiguous queries have multiple meanings, especially for short queries that are 1-3 words long [Sanderson, 2008]. On the other hand, even those queries are unambiguous or have clearly defined meanings. They

¹In this thesis, we adhere to the terms “aspect” and “sub-topic”, and use them interchangeably. Whereas intents or query-intents are the terms used to address information needs from a user’s point of view, aspects or sub-topics refer to pieces of information in documents that satisfy user’s information needs.

might be considered *underspecified* since it is not clear which aspects of such meanings the user is actually interested in [Radlinski and Dumais, 2006; Santos et al., 2011b]. Song et al. [2009] studied and categorised these two types of queries, paid attention in sub-topic retrieval, into *ambiguous query* and *faceted query*¹. They also noted that a user who issues the latter query might look for one of the other aspects by browsing retrieval results or issuing another query.

When an ambiguous query such as “jaguar” is entered to an IR system, diversification stands for addressing all possible senses of the word. In the case of our example, this would result in retrieving at the top of the ranking documents related to many senses of jaguar such as the big cat-like animal, the British luxury car manufacturer (i.e. Jaguar Cars Ltd.), the electric guitar (i.e. Fender Jaguar), the Apple’s operating system (i.e. Mac OS X v10.2), etc. In contrast, when the entered query is unambiguous, such as “Jaguar XKS-coupé”, effective diversification policies consist in retrieving top-ranked documents that are topically diverse while still addressing the user’s query. In our example, these may be web pages about the launch of the latest Jaguar car model including its features and specifications, others providing expert reviews about test drives, and finally others suggesting the nearest dealers where the model is available.

Additionally, the presence of duplicate or near-duplicate documents in search results is, in general, undesirable because users have to endure examining redundant information repeatedly [Bernstein and Zobel, 2005; Zhai et al., 2003]. Users may at most be interested in the fact that such documents exist but certainly view them unfavourably. As an extreme example, a document that actually contains relevant information may be considered non-relevant if users have already seen other documents containing the same information [Clarke et al., 2008]. Nevertheless, in some particular search tasks, users may intend to find all relevant documents or achieve total recall, such as patent search [Bonino et al., 2010; Joho et al., 2010], legal search [Cormack et al., 2010], medical records search², etc. For such tasks, result diversification may *not* be applicable.

Note that topical diversity is not the only approach to result diversification. The effectiveness of a system and thus the user’s satisfaction may be enhanced if docu-

¹It is also known as an *underspecified* or *broad* query.

²<http://groups.google.com/group/trec-med>

ments are diversified with respect to opinions, sources, media format, etc. Aggregated search is an example research area that investigates the impact of the presentation of search results integrated from different sources (web, image, video, news, blog, tweet, etc.). When considering the diversity retrieval task within this thesis, we focus on the topical or content-based diversity although diversification and evaluation approaches developed here can be applied also to other forms of diversity.

2.4 Evaluation in Information Retrieval

This section surveys well-established experimental methodologies in the information retrieval domain. An overview of information retrieval experimentation is provided in Section 2.4.1. Two main experimental approaches that dominate the research field are introduced in Sections 2.4.1.1 and 2.4.1.2. Next, Section 2.4.2 outlines several evaluation measures employed in ad-hoc retrieval and diversity-based retrieval. The most commonly used evaluation measures for both tasks are introduced in Section 2.4.2.1 and 2.4.2.2, respectively.

2.4.1 Experimental Methodologies

One of the goals of scientific research is to evaluate hypotheses and research questions based on clear and justified assumptions. In IR research, evaluation has had a long tradition since the 1960s, when the earliest large-scale evaluation of search performance was performed [Cleverdon et al., 1966]. Evaluation is an important part of IR research to develop better retrieval systems. One of the main goals is to assess whether a retrieval system *effectively* and *efficiently* responds to user's requests (i.e. queries) for a particular search task or application. Whereas effectiveness indicates the ability of the retrieval system to support users finding the right information, efficiency measures how quickly this process is taken. The majority of IR experiments focus on evaluating the effectiveness of retrieval systems.

System effectiveness can be inferred by measuring user satisfaction with the system, i.e. assessing whether users are satisfied with documents returned in answer to their queries. Therefore, IR theories and models are developed with the fundamental objective of maximising user satisfaction given their queries [van Rijsbergen, 1979].

Considering human-computer interactivity, *interactive* IR focuses on improving the system ability to present search results on a graphical user interface and to tailor the results with respect to a user's changing information needs [Beaulieu and Jones, 1998]. To assess which approach or system performs best in satisfying users, IR evaluation principally relies on experimental methodologies that provide robust and repeatable testing even on large scale experiments. In the following sections, we introduce two evaluation methodologies: system-oriented evaluation and user-centred evaluation.

2.4.1.1 System-Oriented Evaluation

The most common evaluation methodology in IR is *system-oriented* or *traditional laboratory-based* evaluation. Its success is due to the well-established design of evaluation methods that allow systematic and objective comparison between retrieval systems. System-oriented evaluation is based on the early work of Cleverdon et al. [1966], who introduced a *test collection* in controlled settings for the evaluation of computer-based retrieval systems (i.e. no user as part of the experiments). This approach is generally referred to as the *Cranfield*¹ evaluation paradigm [Cleverdon et al., 1966]. In this paradigm, researchers assemble a test collection or *evaluation corpus* [Voorhees, 2005], which consists of documents, queries, and relevance judgements as well as the measurements of precision and recall ratios. The Cranfield evaluation constitutes the empirical research tradition of development and testing of IR systems. The emphasis in this research tradition is on controlled laboratory tests. All relevance judgements are pre-defined and all variables are fully controlled. By controlling these experimental parameters, researchers can draw a conclusion from the outputs of retrieval systems. The main concept of the Cranfield paradigm is as follows.

- Set up a collection of *documents*. The purpose of a document collection is to provide a common test bed for evaluation that enables fair comparison between different approaches.
- Create a test suite of information needs, which are usually a collection of *queries* (also known as topics). This sometimes comes together with a brief description

¹Named after the place in the United Kingdom where the Cranfield experiments were conducted.

of the associated information needs. The queries serve as inputs for retrieval systems.

- Gather a set of *relevance judgements* (often called in short *qrel*) for a particular search task. It is generally a binary assessment of either relevant or non-relevant for each query-document pair. This decision is referred to as the *gold standard* or *ground truth* judgement of relevance. The outputs of retrieval systems will be compared with these judgements to quantify the amount of relevant information retrieved by the systems.
- Evaluate the retrieval outputs against the known relevance judgements, in terms of various performance measures associated with the search task. If two or more systems are considered, then they should be compared statistically.

The Cranfield paradigm has become a standard approach for IR evaluation and has been adopted in several evaluation campaigns such as *Text REtrieval Conference* (TREC¹) [Voorhees and Harman, 2005], *Cross-Language Evaluation Forum* (CLEF²) [Gey et al., 2005]. Although the Cranfield collection allows precise quantitative measures of information retrieval effectiveness, it is nowadays deemed too small for comprehensive experiments. This is due to the difficulty of creating a large test collection since the original Cranfield experiments considered complete relevance assessments [Cleverdon, 1991].

With the massive human effort involved, acquiring complete relevance assessments is very expensive and thus not suitable for creating large-scale test collections, e.g. web corpus. As a result, in the evaluation of TREC, relevance assessments may be incomplete, i.e. not all the documents in the collection have been judged with respect to their relevance to all the queries. Spärck-Jones and van Rijsbergen [1975] proposed to create the assessment lists from subsets of the actual collection. This approach is referred to as the *pooling* technique, by which relevance assessments are only performed

¹An on-going series of research workshops, which focus on a list of different IR-related research areas (called tracks) and have been sponsored and organised by the U.S. National Institute of Standards and Technology (NIST). See <http://trec.nist.gov/>.

²A European workshop that aims to promote research in multilingual (mainly European languages) information access. See <http://clef.isti.cnr.it/>. It also included a co-organised evaluation forum (*ImageCLEF*) for the research in cross-language annotation and retrieval of images. See <http://www.imageclef.org/>.

on a “pool” of documents. This pool is usually created by merging results for the top r documents, returned in response to each query from multiple retrieval systems, or typically participating systems at TREC. Assessors then provide relevance judgements only for those documents contained in the pool. It is interesting to note that if many retrieval systems contribute to the pool, the relevance assessments are probably not to be biased towards any particular systems. Besides, a test collection, which includes these relevance assessments as well as documents and queries, can be employed for a reliable evaluation of other retrieval systems which do not contribute to the pool. [Sanderson and Joho \[2004\]](#) evaluated various other approaches to create a test collection without pooling. However, none of the evaluated assessment approaches results in a complete list containing all relevant documents of the collection.

2.4.1.2 User-Centred Evaluation

The Cranfield paradigm treats information needs as a static concept entirely defined by search queries. This implies the assumption that the changing of information needs is disregarded and confined to the queries alone. In fact, information needs change as users encounter new information from search results and these vary individually for each user [[Ingwersen and Järvelin, 2005](#)]. [Robertson and Hancock-Beaulieu \[1992\]](#) argued that system-oriented evaluation is not suitable for assessing interactive IR systems. This is because the controlled evaluation environment ignores cognitive and behavioural features, associated with human decision on relevance assessments. In interactive IR, relevance is considered to be dynamic and evolves over time according to a cognitive state that forms an information need [[Borlund, 2003b](#); [Ruthven, 2005](#)]. Furthermore, users’ relevance assessments are subjective in the sense of either intellectual topicality, pertinence or situational relevance [[Ingwersen, 1996](#)]. The focus of interactive IR evaluation is thus on the user with respect to the system’s overall design and development, including user interfaces, retrieval strategies, feedback mechanisms, etc. Therefore, user’s information use, retrieval, and searching behaviour with the objective of obtaining realistic results are required to be taken into consideration. To inform us about the effectiveness of an interactive IR system, we need to measure how well the system is capable of predicting the relevance of documents (algorithmic relevance) and how well the same documents, deemed topically relevant by the system, are actually

relevant to the user in a particular context (subjective, situational relevance). Consequently, the evaluation of interactive IR systems has to include the user's interactive information searching and retrieval processes [Borlund and Ingwersen, 1997].

Inspired by human-computer interaction and psychology, an alternative evaluation paradigm is *user-centred* or *task-oriented* evaluation. It considers user's natural interaction, subjective perception, and relevance assessment behaviour in the seeking and retrieval processes. Borlund [2003a] introduced this evaluation paradigm as a model for quantifying the effectiveness of interactive IR systems. She argued that interactive IR systems should be evaluated under realistic conditions, i.e. the evaluation procedure should model actual information seeking tasks. Therefore, she suggested to include users as test subjects of IR systems in the experiments. In her model, user's perception and behaviour are the centre of the evaluation instead of system performance measured by common evaluation metrics, e.g. precision and recall. The key idea is to extend the controlled computer-based experiments to the context of realistic search scenarios (by the use of simulated work task situations) while all variables and research situations still remain under control. A simulated work task is a short "cover story" that describes a situation where a certain information need requires the use of an IR system. Users thus have to perform search so as to find information that satisfies their needs.

According to Shadish et al. [2001], the nature of human behaviour is one of the problems affecting a user study. This, in particular, results from natural learning aptitude, by which humans can learn how to handle a system and solve a task. Thereby human behaviour in one condition will influence their behaviour in another. In other words, results of subsequent experiments most likely will be better than the results of earlier experiments. To neutralise the effect of learning, a Graeco-Latin square¹ should be applied to control the variation of blocking factors (i.e. controlled variables) in an experiment. Consider the following example where a researcher will evaluate the effectiveness of two interactive IR systems S_1 and S_2 . Assuming two simulated work tasks T_1 and T_2 , a user has to carry them out once using either system. Here, the user is the primary factor, whereas system and task are blocking factors which will be paired together to form a block of experimental session. To obtain different orders of pairs according to Graeco-Latin square design, blocking factors are assigned randomly

¹A Graeco-Latin square is formed by merging two *orthogonal* Latin square of an $n \times m$ arrangement over two sets of blocking factors, e.g. systems and tasks.

2.4 Evaluation in Information Retrieval

Table 2.1: Graeco-Latin square design of two systems and two tasks. Each row represents an order of system-task pairs, assigned to users to perform in an experiment.

Users	two factors rotation	
	Block 1	Block 2
$U1$	$S1, T1$	$S2, T2$
$U2$	$S1, T2$	$S2, T1$
$U3$	$S2, T1$	$S1, T2$
$U4$	$S2, T2$	$S1, T1$

to rows and columns in a table, with each factor once per row and once per column, and no two blocks contain the same ordered pair. Therefore, all users ($U1, \dots, U4$) will perform search on all evaluating systems and in all given work task situations, but in different orders of unique randomised pairs of system and task. Note that the number of systems and tasks should be concordant so that every system will be used equally by each user. By doing this, the number of tasks should be equal to or N times higher than the number of systems, where N is positive integer.

Table 2.1 shows a Graeco-Latin square design of an above example, in which each user performs two search tasks using two systems in an experiments. As can be seen, user $U1$, for example, start by using system $S1$ for task $T1$, and then use system $S2$ for task $T2$.

To investigate system performance and human behaviours on interaction with the system, alternative methods are required to collect quantitative and qualitative data¹. The analysis of these data can indicate the system's effectiveness based on user's feedback and satisfaction with the systems. Common methods used to gather such data in user-centred evaluation are:

- usage log files analysis;
- questionnaires, consisting of open- and closed-questions;
- user interviews; and

¹Quantitative data are any data in numerical form, e.g. ratings, statistics, percentages, etc. These data are then analysed statistically and the emerging patterns of findings are interpreted. Qualitative data, on the other hand, are those collected by open questions and interviews from participants. They are usually in the form of word data, such as meticulous description and explanation regarding user's experience and perception in an experiment.

- video-based observation, etc.

2.4.2 Measures of Retrieval Effectiveness

Once an IR system has been designed and developed, the system's overall performance should be evaluated to ensure whether it performs as desired or not. Many measures of retrieval effectiveness have been proposed, but most common ones are based on a similar principle, i.e. quantifying the relevance of retrieved documents. A simple but effective approach for evaluating the systems is on the basis of binary relevance, measuring how many relevant documents have been retrieved and how many relevant documents have been missed. For instance, *precision* measures the ratio of the retrieved relevant documents over all retrieved documents, and *recall* measures the ratio of retrieved relevant documents over all the possible relevant documents for a query. As ranking plays an essential role in IR, many effectiveness measures are rank-dependent, i.e. the utility of a relevant document to the overall user satisfaction is weighted according to the rank position at which the document is retrieved. Although in some measures the rank position of a document is not included, the weight of the utility of relevant documents can be derived from different utility models such as a user's browsing model, a document utility model, a utility accumulation model, etc. [Carterette, 2011].

To compare the performance of different IR systems, evaluation measures should be defined in accordance with the objective of the retrieval task. For example, the measures for the ad-hoc retrieval could simply consider a query and relevance assessments made with respect to the query (i.e. a user submits a query and judge received documents according to a query, representing his single information need), whereas the diversity-based retrieval must comprise a query, a set of sub-topics, and relevance assessments made with respect to each sub-topic (i.e. a user specifies his multiple information needs through a single query. He will be satisfied if a system retrieves all documents covering all his needs). The goal of the task and its user model are reflected in the measures that are used to evaluate systems. The following will provide an overview of evaluation measures employed in ad-hoc retrieval and diversity-based retrieval.

2.4.2.1 Evaluating Ad-hoc Retrieval

– *Set-based Measures.*

When the rank of the retrieved documents is trivial, retrieval effectiveness is most commonly evaluated in terms of the following measures.

Precision. Precision refers to the fraction of retrieved documents that are relevant.

$$\text{Precision} = \frac{\#(\text{relevant documents retrieved})}{\#(\text{retrieved documents})} \quad (2.7)$$

Recall. Recall refers to the fraction of relevant documents that are retrieved.

$$\text{Recall} = \frac{\#(\text{relevant documents retrieved})}{\#(\text{relevant documents})} \quad (2.8)$$

The concepts of precision and recall were first introduced by [Kent et al. \[1954\]](#) and analysed in the IR community [[Cleverdon, 1972](#); [Raghavan et al., 1989](#); [Salton, 1971](#)]. Precision and recall are the common measures used consistently throughout the ad-hoc retrieval task. Alternative measures are also adopted. For example, precision at a specific ranking position r , i.e. $\text{Precision}@r$, is useful to assess the system performance achieved after retrieving r documents.

F-measure. Neither precision nor recall alone provide a complete view of a retrieval system's effectiveness. It is advantageous to have two numbers of precision and recall in that one is more important than the other in many circumstances. Therefore, it is not trivial to increase one at the expense of the other¹. An ideal retrieval system should achieve an appropriate balance between precision and recall, tuning them with respect to the objective of retrieval task. F-measure addresses this issue by trading off precision against recall through a parameter $\beta \in [0, \infty]$ [[van Rijsbergen, 1979](#)].

¹Precision and recall have an inverse relationship [[Baeza-Yates and Ribeiro-Neto, 1999](#); [Manning et al., 2008](#)]. Precision falls whilst recall increases as the number of retrieved documents increases.

$$F_{\beta} = \frac{(\beta^2 + 1) \cdot \text{Precision} \cdot \text{Recall}}{\beta^2 \cdot \text{Precision} + \text{Recall}} \quad (2.9)$$

A default value of β is 1. This leads the F-measure, $F_{\beta=1}$, to become the *harmonic mean* of precision and recall.

– *Ranked Retrieval Measures*

Most of the modern retrieval systems produce a ranked list of documents so that users are more likely to encounter relevant documents at the top of the list. By far the most well-accepted browsing model is that of a user scanning down a ranked list of documents one-by-one and stopping at some rank r . Correspondingly, most ranked retrieval measures employ this model to estimate the utility of relevant documents, i.e. pay more attention to documents at the higher ranks. We describe some of the commonly used measures below.

Mean Average Precision (MAP). Average precision (AP) refers to the average of the precision values obtained after each relevant document is retrieved, within the top k documents¹ [Voorhees and Harman, 2005]. This value of AP is then averaged over the set of queries to obtain the mean average precision (MAP), i.e.

$$\text{MAP} = \frac{1}{|Q|} \sum_{i: q_i \in Q} \frac{1}{|R_{i,k}|} \sum_{j: d_j \in R_{i,k}} \text{Precision}@j \quad (2.10)$$

where q_i refers to the query in the set of queries Q , $R_{i,k}$ refers to the set of relevant documents for the i -th query from the top result until the document at rank k ², and d_j is the *relevant* document placed at rank j . MAP provides a succinct summary of the effectiveness of an IR system over all queries in a test collection.

¹Note that in this measure we use k instead of r for a rank position. This is because of avoiding the confusion with R that refers to the set of relevant documents.

²In TREC, a commonly used value of k is 1000, which is, in general, equal to the total number of documents retrieved by a system.

Mean Reciprocal Rank (MRR). Reciprocal rank (RR) corresponds to the inverse of the rank position at which the first relevant document appears. For a set of queries Q , the mean reciprocal rank (MRR) is the average of the RR values for each query:

$$\text{MRR} = \frac{1}{|Q|} \sum_{i=1}^{|Q|} \frac{1}{\text{RR}(q_i)} = \frac{1}{|Q|} \sum_{i=1}^{|Q|} \frac{1}{\text{rank-first-doc-relevant}(q_i)} \quad (2.11)$$

where $|Q|$ corresponds to the number of queries in Q .

Normalised Discounted Cumulative Gain (nDCG). nDCG is a popular effectiveness measure that considers both graded relevance assessments and document utility modelled through the position-based discount function [Järvelin and Kekäläinen, 2002]. The intuition underlying nDCG is that a user browses a ranked list in a top-down manner and is less likely to examine lower-ranked documents. This fact is incorporated by the cumulative gain, i.e. the gain accrued from the graded relevance of documents (judged by the user). The gain of each relevant document is however discounted based on how likely the user will examine the documents. The discounted cumulative gain (DCG) at rank r is defined as follows:

$$\text{DCG}@r = \sum_{i=1}^r \frac{J(d_i, q)}{\log_2(1 + i)} \quad (2.12)$$

where $J(d_i, q)$ is the relevance judgement of the i -th document given a query q in the ranked list, and the logarithmic denominator is the discount factor based on the rank positions of documents. nDCG is obtained by normalising this score by the DCG score of the ideal ranked list.

Expected Reciprocal Rank (ERR). ERR is a recently proposed measure, which assumes the (expected) utility of documents based on a cascade model of user browsing [Chapelle et al., 2009]. In the cascade model, a user views ranked documents from top to bottom and for each document, the user has a certain probability of being satisfied. However, the *utility* of a currently viewed document depends on the probability

that the user is *not* satisfied with documents having been viewed previously. To formalise this, let $P(R_r)$ ¹ denote the relevance probability of a document at rank r , and let $\prod_{i=1}^{r-1} (1 - P(R_i))$ denote the probability that the user is not satisfied with documents from ranks 1 to $r - 1$. Then, ERR is defined based on the expected probability that the user is finally satisfied at rank r :

$$\text{ERR}@r = \sum_{r=1}^n \frac{1}{r} \prod_{i=1}^{r-1} (1 - P(R_i)) P(R_r) \quad (2.13)$$

where $1/r$ is the utility function based on ranked positions. To obtain the probability of relevance $P(R_i)$, Chapelle et al. [2009] suggested to convert relevance grades by:

$$P(R_i) \approx R(g_i) = \frac{2^g - 1}{2^{g_{\max}}}, \quad g \in \{0, \dots, g_{\max}\} \quad (2.14)$$

where g is a relevance grade, e.g. $0 \leq g \leq 4$ when a five-point scale is used.

2.4.2.2 Evaluating Diversity-Based Retrieval

The evaluation of diversity-based retrieval has attracted increasing interest from the research community. An example of this can be found in the evaluation campaign of TREC 2009–11 Web Diversity tracks, investigating and evaluating the performance of systems that aim to promote diversity in the search results. As specified in the TREC guideline², retrieval systems designed for the diversity task should return relevant documents that, taken together, provide a complete coverage for a query while avoiding excessive redundancy. Within this context, various effectiveness measures have been proposed such as sub-topic recall (s-recall), sub-topic mean reciprocal rank (s-mrr), α -nDCG, Intent-Aware measures (e.g. *MAP-IA*, *ERR-IA*). In this section, we provide an overview of diversity-based evaluation measures. Note that the thorough discussion of three measures (i.e. s-recall, s-mrr, α -nDCG) mainly used in this thesis are given

¹This relevance probability is in fact the probability that a user is satisfied with the r -th document. However, we shorten $P(R|d_r)$ to $P(R_r)$.

²See <http://plg.uwaterloo.ca/~trecweb/2010.html> guidelines or Appendix D.

in Chapter 7, where we argue the problem of current measure setting and propose an approach to overcome this problem.

– Set-based Measures

Sub-topic Recall. Zhai et al. [2003] proposed to measure diversity in terms of the coverage of sub-topics of a given query. Specifically, s-recall was defined as the fraction of sub-topics covered by documents up to a given rank r :

$$\text{s-recall}@r = \frac{|\cup_{i=1}^r \text{sub-topic}(d_i)|}{|S|} \quad (2.15)$$

where $\text{sub-topic}(d_i)$ denotes the sub-topics covered by the i -th document, and S denotes the set of all sub-topics relevant to the given query. Intuitively, the greater number of sub-topics covered by the top r documents, the more effective the system.

– Ranked Retrieval Measures

Sub-topic Mean Reciprocal Rank (s-mrr). Zhai et al. [2003] also suggested an extension of the traditional mean reciprocal rank measure for the evaluation of diversity-based retrieval task. Similar to the notion of s-recall, s-mrr is defined as the inverse of the rank at which a complete coverage of sub-topics is achieved:

$$\text{s-mrr}@100\% = \frac{1}{\text{rank-first-complete-coverage}(q)} \quad (2.16)$$

The measure may be further adapted to assess partial sub-topic coverage (e.g. 25%, 50%, 75%, etc.) [Chen and Karger, 2006; Wang and Zhu, 2009; Zuccon and Azzopardi, 2010]. For instance, we define s-mrr at 50% coverage (denoted s-mrr@50%) as the inverse of the smallest rank position at which at least a half of all possible relevant sub-topics have been covered by documents in the ranking.

Intent-Aware Measures (IA). A family of IA-measures have been recently proposed by [Agrawal et al. \[2009\]](#). The notion of IA-measures is that each sub-topic s (query-intent) has a certain probability of belonging to a query q , i.e. $P(s|q)$ ¹, and each document contains information addressing such sub-topics. By incorporating a probability distribution of this, we can compute an evaluation measure for each sub-topic separately. This sub-topic dependent measure then is aggregated by averaging its value, weighted by its sub-topic probability. Following this approach, several common ad-hoc retrieval measures (e.g. MAP, MRR, nDCG, ERR, etc.) can be applied. For instance, intent-aware mean average precision (MAP-IA) can be formally defined as:

$$\text{MAP-IA} = \sum_{s=1}^{|S|} P(s|q) \text{MAP}_s \quad (2.17)$$

where $P(s|q)$ is the likelihood of sub-topic s given query q , and MAP_s is the value of mean average precision obtained when considering documents relevant to sub-topic s . Using the same idea of averaging over sub-topics, [Agrawal et al. \[2009\]](#) extended nDCG to obtain its intent-aware version, nDCG-IA:

$$\text{nDCG-IA} = \sum_{s=1}^{|S|} P(s|q) \text{nDCG}_s \quad (2.18)$$

Similarly, [Chapelle et al. \[2009\]](#) framed their ERR measure for diversity-based retrieval task, obtaining ERR-IA:

$$\text{ERR-IA} = \sum_{s=1}^{|S|} P(s|q) \text{ERR}_s \quad (2.19)$$

It has been argued that the IA-measures tend to give no importance to the retrieval of documents relevant to low-weighted intents (i.e. with small $P(s|q)$) [Sakai et al. \[2010\]](#). As a result, this may in general conflict with the specific goal of the diversity retrieval task (i.e. promote a system that produce a high coverage of sub-topics).

¹In other words, $P(s|q)$ is the probability that a given query q can be interpreted as a sub-topic s . If there is a known probability distribution of all the sub-topics for a query q , then $\sum_{s=1}^{|S|} P(s|q) = 1$.

Novelty-Biased Discounted Cumulative Gain (α -nDCG). [Clarke et al. \[2008\]](#) proposed α -nDCG, a nugget-based variation of nDCG [[Järvelin and Kekäläinen, 2002](#)], for evaluating diversified search results. The main idea of this measure is to compute the gain of each document in terms of information *nuggets*, which are ultimately interpreted as sub-topics or query-intents. Each subsequent presentation of the same sub-topic leads to a discounting return to reflect the decreased value provided by redundant information. That is, the gain for relevant documents addressing *novel* sub-topics is not discounted. In particular, α -nDCG defines the gain of the document placed at rank r as follows:

$$NG(q, r) = \sum_{s=1}^{|S|} J(d_r, s)(1 - \alpha)^{D_{s,r-1}} \quad (2.20)$$

where $NG(q, r)$ is a novelty-biased gain of the r -th document given query q , $J(d_r, s)$ indicates whether document d_r contains sub-topic s , and $D_{s,r-1}$ is the number of times sub-topic s appeared in documents up to rank $r - 1$. The free parameter α has been suggested to control how much diversity is rewarded over relevance, with a common setting of $\alpha = 0.5$ [[Clarke et al., 2008](#)]. The total gain up to rank r is then computed as follows:

$$DCNG(r) = \sum_{i=1}^r \frac{NG(q, i)}{\log_2(1 + i)} \quad (2.21)$$

where $DCNG(r)$ is a discount cumulative novelty-biased gain obtained from documents up to rank r . α -nDCG can be derived by normalising this DCNG score by that of the ideal ranking. Although α -nDCG has become a de-facto measure for the evaluation of diversity-based retrieval, this thesis will show that in some circumstances α -nDCG does not behave as specified in the goal of diversity track. In [Chapter 7](#), we will present and discuss the issue, and also provide the solution to overcome this problem.

Part II

Need for Diversity in Exploratory Search

Chapter 3

Diversity-Based Recommendations for Image Browsing

3.1 Introduction

In Part II of this thesis, we investigate the benefits of diversity in search results through a user study, which considers two interactive information retrieval systems. These systems are compared in terms of user preferences and effectiveness in supporting users, who are engaged in a specific class of information-seeking activities. This class stems from the uncertainty of a user's information need [Ingwersen, 1992] and is referred to as *exploratory search* [Marchionini, 2006; White and Roth, 2009]. Exploratory search is motivated from the fact that the information need is often ambiguous or ill-defined. The task itself is generally concerned with information exploration, as carried out by searchers who are:

- unfamiliar with a certain domain of information (i.e. need to learn about the topic to understand how to find answers);
- uncertain about the terminology used by search systems (i.e. do not know what types of documents are present in collections and how they are represented);
- unsure about the way to achieve their goal¹ (i.e. unsure about what queries/actions they should take so as to obtain the information they are seeking); and/or

¹Goal in this context means the object of human ambition or effort that can fulfil their information need.

- unsure, even, about their goal in the first place (i.e. their underlying information needs are initially vague, “I do not know what I am looking for, but I will know when I find it” [Ter Hofstede et al., 1996]).

Exploratory search is a specialisation of information-seeking activities, where searchers attempt to obtain relevant information through a combination of querying and collection browsing [White et al., 2006]. Due to the lack of prior domain knowledge about the document collection or search topic, exploratory searchers may have no clear steps towards finding the required information or they may find it difficult to formulate well-specified queries. Therefore, a common search strategy often employed by the searchers is to start with a *broad* or *tentative* query for obtaining near-relevant information. They then navigate through the retrieved information, which helps them clarify their search goal [White et al., 2008]. In other words, exploratory searchers do not adhere to typical search strategies, i.e. entering a carefully planned series of queries. They instead employ browsing strategies, e.g. on-the-fly selection, in order to explore information for a better understanding of the search domain. During browsing activities, the obtained information provides searchers with cues about the next steps, which will eventually lead them to achieve their goal.

To this end, the research community has strived to support searchers engaged in exploratory search by developing alternative visualisations and user interfaces. This line of research brings together the work in human-computer interaction (HCI) and information retrieval (IR). Rather than framing the search problem as matching queries and documents for ranking purposes, interactive IR considers HCI principles for devising strategies and tools to involve humans more actively into the search process. Therefore, an important part of the research in interactive IR is directed towards creating *interactive* user interfaces, continuously engaging people in the information-seeking process. Search systems designed for exploratory conditions should legitimise browsing strategies (e.g. selection, navigation, and trial-error tactics), which in turn facilitate exploration for knowledge acquisition and information discovery.

Furthermore, as reflected by models of information retrieval interactions, such as the cognitive model of Ingwersen [1996] and Ingwersen and Järvelin [2005], search tasks are part of a larger context and are often regarded as *complex work tasks*, in

which searchers may be required to carry out multiple related subtasks. For example, in decision making tasks searchers may consider and explore multiple solutions before settling on a single final solution. Other tasks (e.g. writing a school report, making a work presentation, or looking for a new job) may involve searching many different aspects of a single search topic so that searchers can learn as much related information as possible. To support the full range of information-seeking activities, exploratory search systems should provide an overview of possible search *aspects*¹, so that searchers can develop their understanding of a given search domain and carry out complex work tasks that may be infeasible with existing systems.

As part of the efforts to develop an interactive IR system that support exploratory search, our work at the University of Glasgow has been oriented towards the development of a graph-based adaptive browsing system, called *Ostensive Browsers* (OB), which can be applied to a variety of image datasets [Campbell and van Rijsbergen, 1996; Urban et al., 2006]. OB aims to facilitate the exploration of relationships amongst data, by incorporating an adaptive query learning scheme based on implicit user feedback. OB displays such relationships as a graph of user browsing trails, which serves as an alternative to existing search (i.e. querying) for content navigation. With the capability of content-assisted browsing, OB allows users to narrow down their broad search domain and to develop an understanding of their uncertain search goal.

Nevertheless, the original OB system was developed with a single browsing space (like old-fashioned web browsers with a single tabbed view), which supports the exploration of a single search aspect. As a result, it does not support complex work tasks that may be composed of multiple aspects or subtasks. In order to support this, we aim to further develop the OB system providing searchers with *diverse recommendations* and *aspectual browsing interfaces*. The extended system is called *Ostensive Browser Plus* (OBP). This system allows users to discover possibly related aspects through diverse recommendations. At the same time, users are able to define the discovered aspects for further search and categorisation through aspectual interfaces. The system as a

¹The definition of aspect as used in the TREC interactive track is related to the definition of *aspect* or *sub-topic* used in this thesis. The interactive track in TREC-6,7,8 [Hersh and Over, 2000; Over, 1998, 1999] defines an aspect as “roughly one of many possible answers to a question which the topic in effect posed”. The task of interactive searchers is to find documents, which, taken together, cover as many different aspects of the topic as possible.

whole is refined and improved to completely facilitate the exploratory search activities associated with uncertain information needs.

3.1.1 Goal and Plan of the Chapter

In this chapter, we aim to investigate whether users of interactive IR systems benefit from diversity in search results, in particular, when their information need is related to multiple aspects. The focus of this chapter is on the category of extrinsic diversity, as discussed in section 1.1.2. It has been argued in fact that there are a number of situations, and for example exploratory search, where users may intend to find the documents that, taken together, cover all different aspects associated with fulfilling a work task. In such scenarios, it has been assumed that users would prefer systems that provide diverse results over traditional systems [Clarke et al., 2008]. Extensive research had been carried out to develop approaches and systems that promote diversity in search results (e.g. [Chen and Karger, 2006; Zhai and Lafferty, 2006]). However, this proposition has not been studied from the user’s perspective. In order to investigate this issue, this chapter introduces a user study, conducted under exploratory conditions, e.g. using complex work tasks that require users searching on multiple aspects and using search systems that are designed for information exploration.

The goal of the study is to observe and analyse user preferences and effectiveness of systems in supporting work tasks and the amount of preformed interactions. The two interactive IR systems compared in the experiment are:

- 1) the Ostensive Browser system (without diversity feature); and
- 2) the Ostensive Browser Plus system that provides diverse recommendations.

In addition to examine the need for result diversity, we propose that *implicit relevance feedback* can be employed to provide relevant recommendation. Following this approach, users will be able to explore a collection to a greater extent and discover more aspects of a search topic which they may not have considered before. This is, in particular when implicit relevance feedback is used to recommend documents related to their search activities. The recommender system within the OBP system is based specifically upon the graph-based implicit feedback mined from user browsing trails [Leelanupab et al., 2009c].

The rest of this chapter is organised as follows. The next section surveys related work in this area. Then, Section 3.3 outlines experimental systems and introduces our approach to diverse recommendations based on image browsing. The chapter concludes in Section 3.4, where we summarise our recommendation approach and lead to the user experiment conducted in Chapter 4.

3.2 Background and Related Work

3.2.1 Exploratory Search

Exploratory search [Hearst, 2000; Marchionini, 2006; White et al., 2008] is an emerging area of information retrieval research, which focuses on the information-seeking problem that occurs when the information need is ill-defined in the searcher's mind. Searchers must learn whilst searching so that they can improve their understanding and clarify their need. Browsing becomes an alternative search strategy employed to obtain information, which will eventually fulfil their clarified need. Here, we focus on exploratory search, a concept that cover such an information-seeking context. Marchionini [2006] characterised exploratory search as follows:

Exploratory search can be used to describe an information-seeking problem context that is open-ended, persistent, and multi-faceted; and to describe information-seeking processes that are opportunistic, iterative, and multi-tactical. In the first sense, exploratory search is commonly used in scientific discovery, learning, and decision making contexts. In the second sense, exploratory tactics are used in all manner of information seeking and reflect seeker preferences and experience as much as the goal.

Exploratory search is thus a specialised form of information seeking with respect to the problem context and search strategies used. In many ways, exploratory search is as much about the journey through information space as the destination. The destination can be viewed as the relevant document with the sought answer whereas the answer may not be immediately obvious if one's knowledge regarding the search domain is unclear. In exploratory search, this may only emerge after the analysis of information gathered during one's journey. Searchers in exploratory search, therefore,

require systems that support their specific activities, allowing them to explore information for knowledge acquisition towards higher-level learning objectives [White and Roth, 2009].

The early work of Pirolli et al. [1996] is an example of a system designed to support search result exploration. By using text clustering techniques, their Scatter/Gather system presents users with summaries of the contents of clusters of similar documents. Its interface facilitates browsing strategies by allowing users to navigate through summaries at different levels of granularity. Instead of clustering, Hearst [2000] exploited categorical metadata, referred to as a *facet*, to organise search results as a set of disjoint partitions. Many modern commercial websites (e.g. amazon.com¹ and ebay.com²) use this categorical structure, enabling users to browse products based on brand, price, type, etc. In the video retrieval domain, Marchionini [2006] introduced the Relation Browser, which partitions the open video digital library into *slices*, based on the videos' metadata attributes. That system allows users to explore a video collection using attributes such as genre, feature, format, language, etc. In short, these systems feature *automatic* categorisation of facets through available data in document collections.

Villa et al. [2009] introduced a *self-organising* exploratory search system for web search, by which users can classify and organize both their searching process and the results of their searching process: i.e. aspects of a complex search task. The system was developed with an *aspectual* search interface, which contains multiple independent search spaces for users to define respective exposed aspects. Within each search space, an aspect allows users to search and mark relevant web pages. Aspects are visualised as web-browser like tabs, but with more flexibility to move, copy, or organise web pages over different tabs. The system is specifically designed to support complex and exploratory search needs that can be defined and structured by users.

3.2.2 Image Retrieval and Browsing

A challenging problem in image retrieval is the so-called *semantic gap* [Smeulders et al., 2000]: the lack of coincidence between the low-level feature representation of an image and the high-level concepts users associated with an image. This problem can

¹www.amazon.com

²www.ebay.com

be attributed to the uncertainty of image perception regarding the user subjectivity and the context it is regarded with [Rui et al., 1998a]. Different people or the same person in many situations may interpret visual content differently. For instance, one person may focus on an image’s colour feature, whereas another may focus on its texture. Even focusing on the same feature may result in different perceptions of similar images.

As of today, it is still difficult to implement reliable techniques to represent the content of an image. This has also implications on the query formulation process. Users are often unfamiliar with data collections and do not know how the information, in particular multimedia data, is represented in retrieval systems. Hence, they are unsure about which queries should be used to obtain relevant information [Salton and Buckley, 1997]. The problem of query formulation is even more critical in the case of content-based image retrieval, where retrieval techniques is based on extracted image features. Users have difficulty to express their information needs in the form that systems understand or employ for indexing and retrieval.

One approach towards alleviating the query formulation problem is to allow users to search by sample images – or what is called “query-by-visual-example”. This approach uses low-level features available in images, such as colour, texture, shape, orientation, etc., to retrieve visually similar results. By adopting clean interface design, content-based image retrieval systems allow users to navigate through an image collection, using a browsing style approach based on image contents. Rather than using a single image as a query, users can pose for retrieval a series of image queries constructed as a graph of the user’s browsing trails. Although this technique is similar to query expansion, it is different in terms of user engagement in the information-seeking process because users are more actively involved in the search. By clicking on the most relevant images at retrieval time, users can retrieve more relevant results without formulating a new text query. This approach engages the user in a highly interactive experience, putting them in control of the information retrieval process.

Graph-based approaches for image browsing were well studied, for example, in [Herman et al., 2000] and more recently in [Viaud et al., 2008]. Nevertheless, one of the major problems that users face is still the semantic gap; this is due to the low-level visual descriptors used for similarity matching. Rui et al. [1998b] proposed an

interactive relevance feedback¹ technique as a method to bridge the semantic gap. They assumed that high-level concepts can be identified when employing both low-level features and relevance feedback. In order to model high-level concepts, their approach dynamically weights the features with respect to relevance feedback, which users are *explicitly* asked to provide while searching. Users continually rate the relevance of images according to their information needs and perception subjectivity. The results of their study showed that the relevance feedback technique greatly reduces the users' effort in formulating a query, and effectively captures the users' information need. However, relevance feedback techniques based on *explicit* ratings interfere with users' normal searching behaviour. By giving explicit feedback, users are forced to engage in additional activities and thus they are distracted from their search.

Alternatively, [Kelly and Teevan \[2003\]](#) suggested that users' natural interactions with systems (e.g. reading, printing, and selecting documents) can be employed as sources of relevance feedback associated with their underlying information needs. By using *implicit* relevance feedback techniques, information about user interests can be obtained without disturbing the users' workflow. For example, [Seo and Zhang \[2000\]](#) proposed an approach to learn users' preferences by unobtrusively observing their web-browsing behaviours. [Claypool et al. \[2001\]](#) examined the correlation between explicit ratings and several browsing behaviours, such as mouse clicks, scrolling, and time spent on documents. [Maglio et al. \[2000\]](#) suggested to infer attention from capturing users' eye movements. Within the HCI community this has become a widely used technique for gathering implicit feedback, e.g. [[Beymer and Russell, 2005](#); [Buscher et al., 2008](#)].

[Urban et al. \[2006\]](#) introduced a content-based image browsing system, Ostensive Browser, that uses implicit relevance feedback from user's clicks on images. Using this feedback, the system constructs an image graph of user browsing trails and exploits it for expanding queries. Furthermore, the system applies the Ostensive Model of Developing Information Need [[Campbell and van Rijsbergen, 1996](#)] to weight the user feedback according to the time of user interactions (e.g. clicks). [Campbell \[2000\]](#) investigated different relevance weighting profiles for the model: flat, increasing, current,

¹Typical relevance feedback is used for query expansion during short-term modelling of a user's immediate information need, and for user profiling during long-term modelling of a user's persistent interests and preferences.

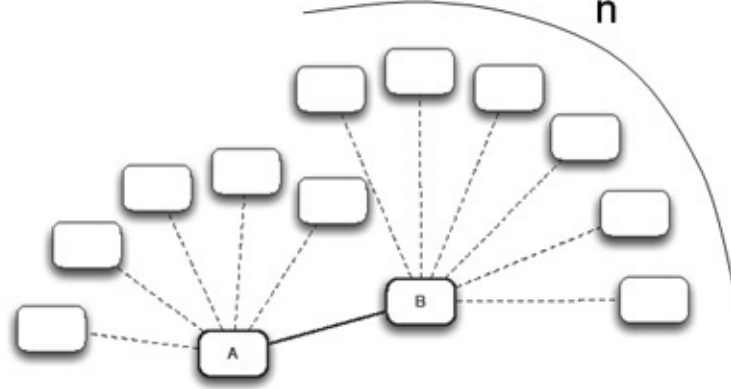


Figure 3.1: Graph-based image browsing

and decreasing profiles. It was shown that the decreasing profile¹ was most effective in tailoring the image results to the user's *current* information need. This is because the user information need and the knowledge of the search domain were gradually developed after subsequent interactions with the system. As a result, the subsequent information in the graph of browsing trails is assumed to be more relevant to the user.

Figure 3.1 illustrates an example of Urban et al.'s approach [Urban et al., 2006]. Given an image *A* as a node in a graph, similar images are shown as leaves of this node. Selecting one of these leaves (i.e. image *B*) will implicitly provide relevance feedback, allowing the system to adaptively retrieve other similar images related to that leaf. Here, a query is based on the path formed by the selected nodes, and referred to as an *ostensive query*. This path represents the user's exploration of information, and taken as a whole is used to build a representation of the immediate information need. In other words, both nodes *A* and *B* are considered for forming the query. The weight of how much each node contributes to the query can be chosen depending on the uncertainty model or the weighting scheme, referred to as *ostensive relevance profiles*. These profiles reflect how relevance (or uncertainty) changes with time (time here being interpreted as the order of selection or the position of nodes in the browsing path). This approach constructs an image graph by taking into account the structural relationships between images based on users' feedback.

¹Lower weighting was given to earlier feedback.



Figure 3.2: Example of similarity-based image recommendations (a) and diversity-based image recommendations (b) [Deselaers et al., 2009].

3.2.3 Recommender Systems

Document recommendation is an alternative personalisation technique¹, exploiting relevance feedback to provide additional relevant documents. The main idea of this technique is to provide users with information that they might be interested and not require extra user interaction. Consequently, users spend less time and search effort on finding relevant information. With recommendations provided by a recommender system, users are presented with lists of documents that are similar to those they have previously searched or viewed.

In many commercial websites (e.g. amazon.com², ebay.com³, jinni.com⁴), recommender systems contribute to better customer experiences and enhance success in meeting customer's needs [Liang et al., 2007]. Users are exposed to relevant items (e.g. products, movies, services) that they might be interested in and not be aware of before. Furthermore, Tam and Ho [2006] found that recommender systems are beneficial to boost cross-selling and increase customer loyalty when recommendations are personalised to individual customers. To increase a chance of retrieving information that interests users, they suggested that recommender systems should include diversity into their recommendations. Deselaers et al. [2009] presented an example of diversity

¹Note that various personalisation techniques have been proposed in the literature. Jameson [2008] reviewed and listed those techniques, including: taking over parts of routine tasks, adapting user interfaces, giving assistance about system use, personalising search results, tailoring information presentation, recommending relevant information, etc.

²www.amazon.com

³www.ebay.com

⁴Online movie search and recommendation service: www.jinni.com.

in product search (see Figure 3.2), where a customer has been recorded for his previous purchase of a certain product, e.g. camcorder, with an exact name of the purchased model. The recommendation functionality of e-commerce sites should not retrieve the set of homogeneous results only of such a camcorder, but instead should suggest the set of diverse results, including related items of interest (camcorder accessories) that might interest the customer. Ziegler et al. [2005] proposed a topic diversification method for recommendations. Their user survey showed that diversity in recommendations improve user satisfaction. Nevertheless, no user study was conducted to investigate the advantages of result diversity in the context of a given work task situation, where users are required to find different aspects of a search topic.

In the case of our research, recommender systems play a role in exposing search aspects that users may use to fulfil a multi-aspect related work task. At the beginning of the search process, users browse through a document (image) collection, aiming to find the desired information regarding one search aspect. Meanwhile, the system suggests a set of related documents, which are diversified to provide other aspects that might be useful to complete the task. Users define the obtained aspects as browsing spaces in aspectual interfaces, where they can browse and mark relevant information with respect to those aspects.

Recommendation algorithms are broadly categorised into collaborative, content-based, and hybrid. Collaborative filtering is considered a social information filtering technique, which involves recommending items based on preferences of users who share similar interests [Schafer et al., 2007]. Unlike collaborative filtering, content-based filtering is a method that determines the relevance of items (e.g. textual documents, images, videos) based on user's own interests and item information [Pazzani and Billsus, 2007]. In the content-based recommendations, user interests are collected from either user's previous feedback (for immediate information need) or user profiles (for persistent interests or preferences). Items are compared with such user interests, and the most similar items are recommended to the user. Hybrid approaches combine both collaborative and content-based filtering to increase recommendation performance [Choeh and Lee, 2008]. Note that our work focuses on the content-based filtering technique since it identifies immediate user interests during user's browsing.

3.3 System Description

3.3.1 System Overview

In order to find out whether users need a system that provides diverse results, a user study is required where participants carry out exploratory work tasks and use experimental systems designed for information exploration. We opt for an image exploratory search system, *Ostensive Browser* [Urban et al., 2006], for the purpose of the study. Although designed to support exploratory search, the original OB system was developed with a single browsing space. As a result, the OB can only deal with a regular (ad-hoc) search task, in which users have to find information about a single aspect. We hence further implement and refine the system to support a broad and complex search task, which may consist of multiple aspects or subtasks.

Similar to the aspectual search system of Villa et al. [2009], we implemented aspectual interfaces to facilitate *user's browsing* in the OB. In order to support users in information exploration, Villa's system includes aspectual interfaces for simple textual search of web pages. To the best of our knowledge, the OB system is the first to include aspectual interfaces for content-based image browsing. Moreover, we develop the diversity-based recommender system, which uses implicit relevance feedback to provide relevant and diverse recommendations. The recommendations are based on a graph of user feedback, as we will explain in details in Section 3.3.3. The improved system is the *Ostensive Browser Plus*, which includes the two main additional features: *diverse recommendations* and *aspectual browsing interfaces*. Further information about the architecture and implementation of OBP system is given in Appendix A.

Figure 3.3 shows an overview of the OBP's components. The upper component is the aspectual browsing interface implemented in a client system. The interface is composed of two types of user interface: traditional text search and content-based browsing. The text search interface provides a common feature, which allows users to enter text queries for searching images. From text search, the results will serve as example images to start content-based browsing (i.e. search by image example). The content-based browsing interface is built around the concept of search aspects, where each aspect contains the following elements:

3.3 System Description

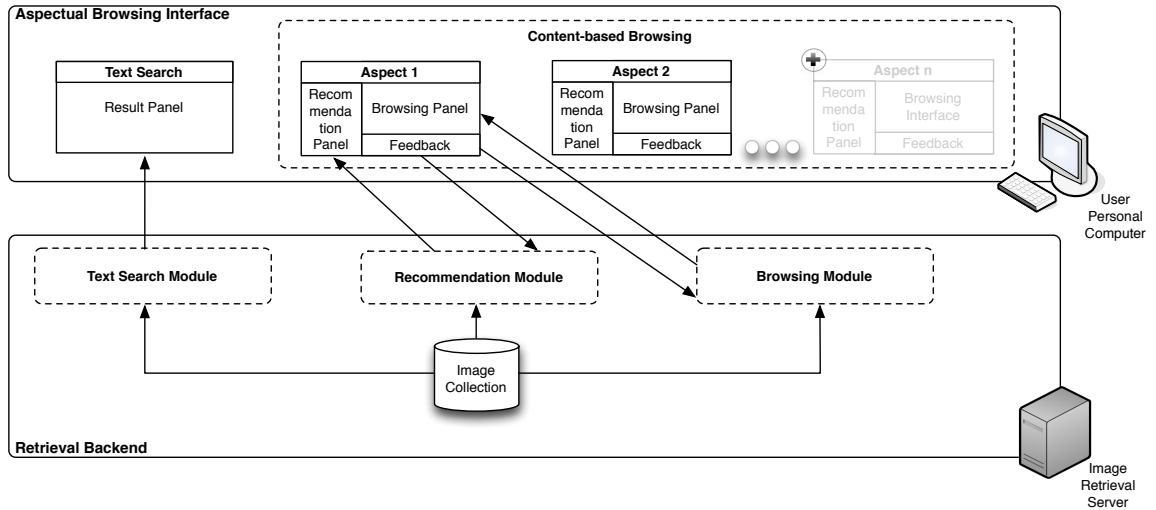


Figure 3.3: The components of the aspectual browsing system with recommendations

- 1) a name, which is by default set to the image title used to start the aspect, but which can be explicitly set by the user when desired;
- 2) a browsing space of OBP, where users explore an image collection by clicking on images executed as queries; and
- 3) a list of recommendations, i.e. the corresponding images which are suggested based on user interactions in the browsing space.

The interface can support as many aspects as defined by users from the obtained images that they consider covering new aspects. It should be noted that an aspect is a self-contained entity, containing all of the above states: each aspect has its own browsing space, recommendations, etc.

In Figure 3.3 the lower component of the architecture is the image retrieval server, consisting of three modules: text search, browsing, and recommendation modules. The text search module performs image retrieval based on text features obtained from image descriptions. The browsing module employs the Ostensive Model of Evolving Information Need [Campbell and van Rijsbergen, 1996] to weight user's feedback according to its time for adapting retrieval results. The intuition underlying this model is

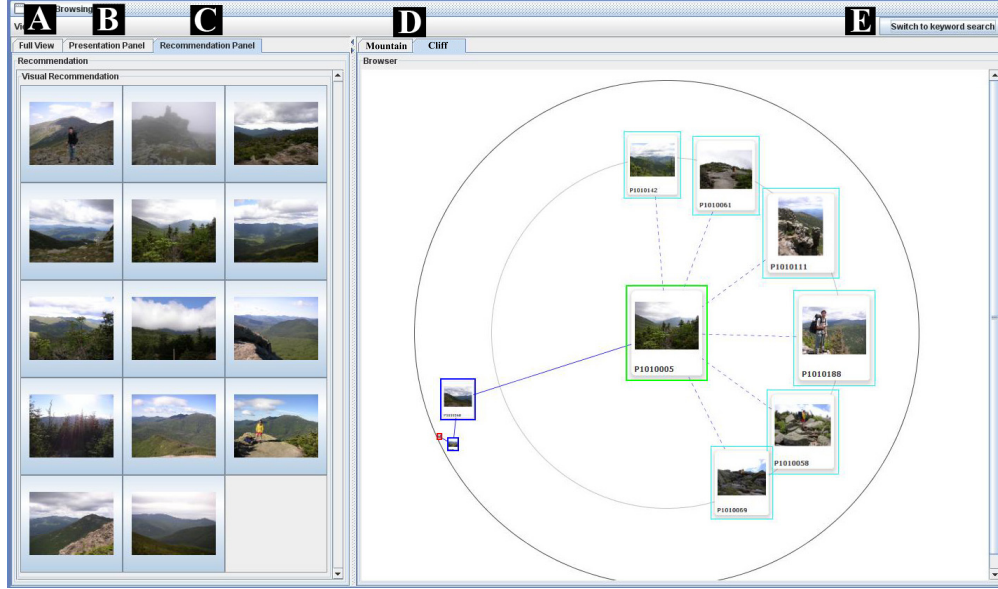


Figure 3.4: Browsing interface of the OBP system

that user's information need is non-static. It develops over the time during the search and is influenced by the documents retrieved. Therefore, this model suggests to modify the weighting of document features provided as feedback based on the iteration in which a user interacts with the corresponding document. We use the decay weighting in the OBP system as [Campbell \[2000\]](#) advised that it is the most effective profile to model developing user interest. Note that we do not focus our research on testing different profiles of the Ostensive Model as this is out of the scope of our study. For more details, the reader is referred to the work of [Campbell \[2000\]](#) for a thorough investigation of other ostensive relevance profiles. For the recommendation module, the system exploits user's feedback to provides a set of diverse results, modified to cover many visual aspects that users may be interested in.

3.3.2 Interface Design

In this section, we illustrate the graphical interfaces of the OBP system. The system fully supports drag and drop operations and consists of two main user interfaces: *browsing interface* (Figure 3.4) and *slide-show interface* (Figure 3.5). The

browsing interface is divided into two main panels. The left panel is composed of: full view tab (A), showing a full size visualisation of the image, accompanied with its textual metadata; presentation tab (B), containing list of relevant lists marked by users; and recommendation tab (C), where users are presented with image recommendations. Note that the presentation tab (B) is designed in consideration of simulated work tasks that users need to perform in an experiment, i.e. making a work presentation of images.

The recommendation tab (C) is allocated to present diverse recommendations and *only* appears in the recommender system, i.e. the OBP system with the diversity feature. A user can choose these recommended images to either start new browsing aspects in the right panel of the interface (D), or to mark as relevant by adding them into the presentation tab (B). Furthermore, browsing aspects can be initiated by selecting images from the results of text search or of other browsing aspects. It is assumed that a user defines new aspects when finding an image covering the corresponding aspects. These aspects are independent and visualised as tabs in the browsing panel (D). Each tab contains a browsing space, in which a user clicks on an image, considered most relevant, to retrieve other similar images for further navigation. In the browsing space, images are linked together by a path representing the user's actions and therefore interest in accessing information. In addition, this query-less interface allows users to jump back and forth between images in one path and branch off into different directions if they realise that they have navigated through a wrong direction. At the top right of the frame, a switching mode button (E) is provided in order to offer the users the option to change search methods between text search and content-based browsing.

Figure 3.5 shows a screenshot of an active presentation tab (B). With the design to simulate search scenarios of complex work tasks given to users, the OBP features a photo slide-show for running presentation. In such scenarios, users are asked to create work presentations of images, taken together, covering multiple aspects of search topics. Users select relevant images and put them into the presentation panel. In this panel, users can modify their presentation by inserting, updating, or deleting the selected images. A click on the play button (1) will start an animated slide-show in a slide-show window (2). In this window, the users can move forward or backward through an size-increased presentation of each image. Moreover, by clicking on a play button (3), they can trigger the automated slide-show where each image will be displayed for one second, followed by the successive image.



Figure 3.5: Slide-show interface of the OBP system

3.3.3 Recommendation Approach

As argued in Section 3.2.2, a graph-based representation of image browsing provides a user with an easy access to image collections. However, Figure 3.1 illustrates a limitation of this presentation technique. Assuming that a search returns m relevant images, only a small set n of these results can be displayed to maintain the usability of the interface. This results in $(m-n)$ potentially relevant images that are not inspected by the user. Besides, this set of neglected images are the results based on users' provided implicit relevance feedback and therefore represents user interest based on their recent browsing interactions. In other words, we exploit implicit relevance feedback extracted from user browsing trails to retrieve a set of $(m-n)$ potentially relevant images as a source to generate recommendation for users.

To create a set of these potentially relevant images, let Q be a set of I ostensive queries used in a browsing aspect or, in other words, is a set of all images selected by users during browsing. q_i is an ostensive query at i -th composed of images in a path that a user selects. img_t is an image within the ostensive query, e.g. q_1 , where t is the time at which the image is selected and used for weighting based on ostensive

Algorithm 1 Creating a set of potentially relevant images as candidates for recommendations

Require: $Q = \{q_1, q_2, q_3, \dots, q_I\}$, a set of ostensive queries q

Require: $q_i = \{img_1, img_2, img_3, \dots, img_i\}$, a set of selected images in a browsing path used as a query by example.

$A_0 = \{\}$

for each $q_i \in Q$ **do**

$A_i = A_{i-1} \cup ORel_{(m-n)}(q_i)$

end for

return $A_i = \{img_x, img_y, img_z, \dots\}$, a set of candidate images accumulated to generate recommendations at the i -th query

relevance profiles. $ORel(q_i)$ is an ostensive retrieval function that retrieve the top m ranked in the result lists where n is the number of images presented to the user and $(m - n)$ is the number of potentially relevant images collected for recommendations. Since we want to provide the recommendations from this $(m - n)$ images, let us define $ORel_{(m-n)}(q_i)$ as a function that returns only the $(m - n)$ images. A_i is a set of candidate images accumulated to generate recommendations at the i -th query within the browsing aspect. The algorithm used to accumulate images is outlined in Algorithm 1.

In order to assist users to carry out complex work tasks, recommender systems should provide recommendations covering multiple aspects of images. In image retrieval domain, the common modality used for image search is text, used in both indexing and retrieval. Although not without its flaws, tags and textual descriptions of photos prove to be reasonable ways to describe and retrieve relevant images. Nevertheless, at the same time text can provide little information about the rich image contents. As the classic quote states “A picture is worth a thousand words”, this is simply because images convey information that words cannot capture, or at least not in any practical setting [Rüger, 2009]. Furthermore, using textual retrieval modality for image search may lack visual diversity. For example, images of the London natural history museum tend to show the same touristic hotspot often taken from the same angle and distance, or the same pictures released by the marketing division of national museum organisation. This absence of visual diversity is crucial since there are several aspects that are not sufficiently covered by the textual annotation. We therefore focus on visual diversity so as to examine the benefits of including it in the results of recommendations. van Leuken et al. [2009] proposed to use lightweight clustering techniques to diversify

results based on several visual features, i.e. colour layout, scalable colour, edge directivity, tamura, etc. In our case, as the purpose is to study the diversity effect and not to propose a new diversification approach, we thus follow their paradigm to diversify results using a clustering technique.

In the recommendation model, we employ a hierarchical agglomerative clustering with the single linkage method to generate recommendations from a set of candidate images A_i . The key idea behind using clustering for image selection is to model a hypothetical set of (visual) aspects, represented by the clusters of images. Images with similar visual appearances will be generally grouped into the same cluster. Furthermore, clustering is performed independently based on different visual features. These features represent different aspects of the images, such as colour, edge, and texture. Three MPEG-7¹ image features are employed in the experiment, i.e. colour layout descriptor (CLD), edge histogram descriptor (EHD), and homogeneous texture descriptor (HTD). Each feature has its own representation and a corresponding similarity matching method. For computing distances between feature vectors, specific similarity functions are employed in accordance with three different features for clustering and retrieval. We describe the basics of low-level signal measurements, termed features, and their particular similarity matchings implemented in OBP in Appendix C. It should be noted that the extraction and similarity metrics of visual features are not the main objectives in this study.

The clustering algorithm generates three dendrograms, which are built by progressively merging the closest cluster until k clusters remain. We assume that each cluster has the potential to reflect different aspects of the user's information need. Hence, recommending representative images from every cluster can provide users with a variety of distinct aspects. As suggested by Urruty et al. [2009], we thus select the medoid² as representative of that cluster since it is assumed that it could be the best representative of the cluster. To avoid overwhelming the user, we set $k = 5$ as a maximum of five selected images from each feature, following the advice of Miller [1956]. A recommendation list can contain a minimum of five and a maximum of 15 images due

¹MPEG-7 is a standard to support a broad a range of applications, devices, or computer codes for describing multimedia content data (e.g., image, audio, and video).

²The element, i.e. document, closest to the centroid of the cluster

to possible intersections amongst these images from different dendrograms. These images are then arranged in a random order to the recommendation list. The diversity of this recommendation list is two-fold: First of all, recommending images from each cluster results in a more diverse image selection of different aspects. This diversity is further extended since the clusters are based on different low-level features. Finally, the random order of the images in the list guarantees that all clusters are treated fairly. Note that a comprehensive review and study of the approaches for result diversifications is presented in Part III; where we categorise the cluster-based approach used in this chapter into the *sub-topic aware paradigm*.

3.4 Summary

In this chapter we have discussed a particular class of information-seeking activities derived from the uncertainty of a user's information need. This class is referred to as exploratory search, where at the beginning of the search the information need of a user is ambiguous or ill-defined. Users have to acquire knowledge whilst searching to improve their understanding of the search domain. To support the users' engagement in exploratory search, we have argued that IR systems should support alternative information-seeking strategies of browsing such as selection, navigation, and trial-error tactics. Users are thus allowed to explore various aspects of coherent information and clarify their search goal. We have discussed that in exploratory search scenarios users may intend to find documents that together cover all different aspects associated with fulfilling a work task. Users would therefore prefer systems that provides diverse results over traditional systems. However, previous studies have not shown if the diverse results are actually beneficial to users of IR systems, in particular when a work task is composed of multiple subtasks.

Aiming to investigate the diversity's advantages from the user's perspective, we have introduced a graph-based adaptive browsing system, called Ostensive Browser Plus. The OBP is refined and implemented to facilitate the exploratory search activities. In addition to content-based image browsing, OBP is featured with diverse recommendations and aspectual browsing interface. Our recommendation approach presents documents that are not only relevant but also diverse in terms of various aspects of image contents. A set of recommended images are mined from implicit rel-

evance feedback that users provided whilst browsing. A hierarchical clustering technique is applied on different visual features to select and diversify recommendations. The aspectual browsing interface is developed around the concept of self-organising exploratory search systems. The interface consists of multiple independent browsing spaces, by which users can classify and organize their searching process and results. The OBP system is specifically designed to support complex and exploratory search needs that can be defined and structured by users.

In the next chapter, we shall examine whether users of retrieval systems benefit from result diversity. We present a user experiment conducted under exploratory conditions. For example, we considered complex work tasks that require users searching on multiple aspects, content-based image retrieval in which users are unsure about how retrieval is processed in terms of low-level features, and the OBP system that is designed for information exploration. The experiment will be evaluated by analysing usage log files and questionnaires from a number of users. By doing this, we can evaluate the systems and their respective performances both quantitatively and qualitatively.

Chapter 4

User-Centred Evaluation of a Diversity-Based Recommender System

4.1 Introduction

The need for IR systems that include diversity in search results have been discussed by the research community; however, no previous study has verified this requirement, considering real user interactions and preferences. To investigate this issue, this chapter is focused on a user study comparing two image retrieval systems. One of them provides a set of documents that together cover different aspects of a coherent topic, indicated by user's implicit relevance feedback. By mining implicit user interaction data, user's intentions toward sought information can be inferred. Thus, IR systems can retrieve relevant documents without requiring any extra effort from the user.

We have proposed recommendation based on user feedback, which is mined from a graph of user browsing trails. This feedback is then used to generate a set of potentially relevant documents. To select documents for recommendation, our approach applies a diversification technique based on clustering. By doing this, recommendation can cover many different aspects that users may be interested in or may use to fulfil a work task that consists of multiple subtasks. Furthermore, clustering is independently performed on different visual features (i.e. colour, edge, and texture) so that recommendation is diversified in terms of different aspects of image contents.

As discussed in Chapter 3, IR systems designed for exploratory search should support browsing strategies. Additionally, in exploratory search tasks (e.g. decision making tasks) users may prefer to consider and survey multiple solutions before settling

on a single final solution. They aim to learn as much related information as possible to be able to carry out the tasks. In the study of this chapter, we opt for the Ostensive Browsing Plus system, a graph-based adaptive browsing system featured with diverse recommendations and aspectual browsing interfaces. All experimental volunteers are asked to perform work tasks, making image presentations for different simulated scenarios. These tasks require the participants to explore different aspects of a search topic. The outcome of the study shows that diverse results through recommendation are beneficial to users when they have a multi-aspect information need.

This chapter is structured as follows. Section 4.2 outlines the experimental plan and research questions investigated in this study. We present the results of the experiment in Section 4.3 and discuss our findings in Section 4.4. Finally, the obtained results and our contributions are summarised in Section 4.5.

4.2 Experiment and Validation

Next we illustrate the experimental methodology based on a user-centred evaluation. We first define the research questions that our study aims to answer. Then, we define the assumptions that will drive the development of the experiments. Finally, we outline the experimental plan so as to ensure that the collected data can adequately answer the research questions.

4.2.1 Research Questions

As discussed before, the OBP system is designed to facilitate exploratory search activities. It allows users to put less effort in formulating queries, to discover more aspects of images, and to find relevant images that fulfil multiple subtasks. In our user study we aim to answer the following research questions.

- **RQ1:** How useful are diverse recommendations from the users' perspective when they are engaged in complex work tasks, composed of multiple aspects or subtasks?
- **RQ2:** Do users discover more aspects of a work task when using the recommender system than the baseline system (i.e. without recommendations)?

- **RQ3:** How do users define such aspects in the systems? What sources return images that cover new aspects i.e. text search, content-based browsing, or diverse recommendations?
- **RQ4:** Can implicit relevance feedback be used for generating recommendations to support users given the search tasks? How effective are the recommendations based on implicit relevance feedback extracted from the users' browsing?

4.2.2 Experimental Assumptions and Scope of the Study

In order to investigate the need for diversity in exploratory search, we define some assumptions that simulate exploratory search scenarios. In the following, we list the experimental assumptions:

- 1) We employ a “simulated work task” situation [[Borlund, 2003a](#)], where a search topic is set into the context. Such context is narrated by a cover story, describing a situation where a certain information need requires the use of an IR system. In our study, we assume various simulated work task situations, each of which requires users to make a work presentation of images. These images together must cover different aspects of the search topic that is suitable for the given task situation.
- 2) We assume that users are uncertain about the terminology that content-based image retrieval systems employ to represent image data and perform retrieval. That is, they do not know how retrieval is processed in terms of low-level features.
- 3) Given a multi-aspect information need stimulated by simulated work tasks, we assume that users require a system to support exploratory search activities. We thus use the search system, i.e. *Ostensive Browser Plus*, which is designed for information exploration.
- 4) Despite limiting the scope of experiments, the above assumptions are necessary to define exploratory search context and thus able to simulate the situations where users require a search system providing diversified search results. We believe that these assumptions can be changed when fewer restrictions are imposed, but the users' need for diversity of information should still remain.

- 5) Furthermore, this experiment focuses on empirically validating the benefits of diversity in image retrieval. Hence, result diversification based on the visual content is valid to the evaluation contexts, systems, and tasks provided to the users.

4.2.3 Plan of Experiments

4.2.3.1 Experimental Design

To answer the research questions in Section 4.2.2, a user evaluation was conducted where participants carried out four different work tasks using our two experimental systems. The two systems are:

- 1) the OB system (no recommendation) – the baseline system (*S1*); and
- 2) the OBP system – the recommender system (*S2*).

Both systems have similar features, such as text search, aspectual browsing interfaces, and animated presentation. However, they differ in that the latter system *S2* features the recommendation functionality. Each participant performed two tasks using the baseline system and two tasks using the recommender system. We adopted a variance of the Graeco-Latin Square design for rotating and counterbalancing systems and search tasks (independent variables). The design rotates the order of systems and tasks undertaken by the participants so as to reduce learning effects, which can affect the outcome of the study (dependant variables). Table 4.1 shows the order of systems and tasks assigned to each user using a Graeco-Latin square rotation.

The experiment started with an individual introductory session, where participants were given an information sheet and demonstrated how to use the two experimental systems. This introduction took approximately five minutes, and was followed by a training session, where each participant was allowed up to ten minutes of interaction and familiarisation with the systems. After training, they were asked to perform four complex work tasks, as defined in task descriptions. For each task, they had a maximum of twenty minutes to carry it out. After two tasks, a five minute break was given to the subjects, as required by the ethical regulations at University of Glasgow.

We investigated the nature of information exploration using six measures: 1) user perception of search experience; 2) the number of clicks performed whilst browsing;

4.2 Experiment and Validation

Table 4.1: The experimental design follows a Graeco-Latin square rotation for systems ($S1-S2$) and tasks ($T1-T4$), involving 24 users ($U1-U24$)

			systems and tasks rotation			
User			Slot 1	Slot 2		Slot 3 Slot 4
$U1$	5 min introduction	10 min training	$S1, T1$	$S2, T2$	5 min break	$S1, T3$ $S2, T4$
$U2$			$S2, T2$	$S1, T3$		$S2, T4$ $S1, T1$
$U3$			$S1, T3$	$S2, T4$		$S1, T1$ $S2, T2$
$U4$			$S2, T4$	$S1, T1$		$S2, T2$ $S1, T3$
...		
$U24$			$S2, T4$	$S1, T1$		$S2, T2$ $S1, T3$

3) the number of textual queries executed; 4) the number of aspects defined; 5) the number of relevant images found; and 6) the distribution of search methods, i.e. text search, browsing, recommendations, returning relevant images and aspects to users. The systems and their respective performances were hence evaluated both qualitatively and quantitatively.

The users' interactions with the system were logged and they were asked to fill out a number of questionnaires. The experiment started with an entry questionnaire, where users were asked to provide personal background and to rate their experience in image retrieval. After each work task, we asked them to fill out a post-task questionnaire, aimed at understanding their opinion about the task and the system that they used to perform that task. Finally, an exit questionnaire was provided where the participants were asked to compare the two systems. All experimental documents are presented in Appendix B.

4.2.3.2 Collection and Data Pre-Processing

For the purpose of this evaluation we employed the photographic collection derived from the CoPhIR¹ collection. The current collection contains 54 million images uploaded to Flickr² by real users. In our study, we used a subset of approximately 20,000 images taken by unique users during 6 months between 1 October 2005 and 31 March 2006. We selected this time period because it covers the highest density of images

¹<http://cophir.isti.cnr.it/>

²<http://www.flickr.com/>

from unique users, and thus is likely to contain many redundant images taken from the same angle and distance by the same user. We therefore can see the benefit of presenting images that are diversified visually. Images are enriched with textual metadata, which are derived from titles, descriptions, and tags given by Flickr users. For text retrieval we used the open source retrieval engine Terrier¹ [Ounis et al., 2007] to index the collection with stop-words removal and stemming. The Okapi BM 25 is used to rank retrieval results for text query as well as image query (i.e. content-based browsing – ostensive query).

For content-based browsing we used three standard MPEG-7 image features such as colour layout, edge histogram, and homogeneous texture. These features were already extracted and included in the CoPhIR collection. For visual similarities, each feature was calculated individually according to its particular similarity metric (see Appendix C). The similarities of different visual features were then normalised and combined using a linear combination with equal weights assigned to all three visual features:

$$\sum_{f=1}^F \frac{1}{F} \times VSim_f(img_q, img_c)$$

where $F = 3$ is the number of total visual features used, and $VSim_f(img_q, img_c)$ is the normalised similarity of visual feature f between an image query img_q and a candidate image to be retrieved img_c .

To merge the similarities from different sources, i.e. textual and (combined) visual features, the *Dempster-Shafer Theory of Evidence Combination* was applied [Jose, 1998]. We subsequently obtained the pairwise similarity values between a single query image and a candidate image to be retrieved. To obtain the final similarity based on the ostensive query, similarities of all images (i.e. nodes) in a path of a user’s browsing trail are weighted using a decreasing profile of ostensive relevance: $w_t = \frac{1}{2^{t-1}}$, where t is the position/time of a node in a path, starting from 1 for the most recently clicked node. A linear combination is then used to combine the similarities of all images:

¹<http://ir.dcs.gla.ac.uk/terrier/>

$$\sum_{t=1}^T w_t \times \text{Sim}(img_{q,t}, img_c)$$

where $img_{q,t}$ is an image query at t -th, T is the total number of images in an ostensive browsing path, and $\text{Sim}(img_{q,t}, img_c)$ is the image similarity merged from textual and visual features using the Dempster-Shafer Theory. Candidate images are ranked in decreasing order of their similarities. The top n images are presented to users as results of browsing and $m - n$ images are collected as sources of potentially relevant images for recommendations. In this study, we defined $n = 6$ and $m = 20$. For further details about the implementations of ostensive browser, we refer interested readers to the work of [Urban et al. \[2006\]](#).

4.2.3.3 Search Tasks

For experimental work tasks, users were given specific task descriptions and allocated time to find images covering many aspects relevant to those tasks. As suggested by [Borlund \[2003a\]](#), each task provides a *simulated work task scenario* and an *indicative request*, so as to help users understand the search context and stimulate their information needs. The simulated work task scenario outlines a contextual description, giving users a goal and purpose of the task to find images. The indicative request provides a guideline or requirement for the images they need to search for.

[Voorhees and Harman \[2005\]](#) argued that at least 24 different tasks are required to gather statistical significant results from such user experiments. Therefore, most datasets such as TREC consist of at least 24 different search tasks. Nevertheless, to study system specific research questions with reasonable cost and effort, a well established approach (e.g. [[Halvey et al., 2009](#); [Hopfgartner et al., 2008](#); [Villa et al., 2008](#)]) is to limit the number of tasks that the users carry out. For this study, we follow this limitation by creating four simulated work tasks for experiment. All four tasks, in general, ask users for images that cover as many aspects of a search topic as possible while three examples of prerequisite aspects are provided for each topic. The topics of the four tasks used in this study were:

- T1: Wild Living Creatures* – find relevant images showing different species of wild animals. The images should cover at least the following aspects: terrestrial animals, aquatic animals, birds, etc.
- T2: Man-made Vehicles* – find relevant images showing different types of vehicles. The images should cover at least the following aspects: car, train, ship, etc.
- T3: Marine Ecology* – find relevant images showing different natural water resources. The images should cover at least the following aspects: headspring, estuary, river, etc.
- T4: Beautiful British Scenery* – find relevant images showing the scenes of different attractive places in rural areas of UK for visits. The images should cover at least the following aspects: cliff, mountain, castle, etc.

To fulfil the work tasks, users are asked to tailor a presentation of images for a specific simulated situation. For example, in task *T1* participants were asked to find different aspects of wild living creatures. The simulated situation was “Imagine you are a graphic designer of an activist organization for wildlife rehabilitation. Your task is to prepare an image presentation on various subjects of the Wildlife Conservation (WLC). The presentation is aimed at calling general awareness for endangered species and preservations of their habitats. You want to create a short presentation about the variety of wild living creatures.” The other simulated work task situations can be found in Appendix B.

4.2.3.4 Participants

24 participants took part in the user study. The participants were mostly postgraduate students and research assistants. The group consisted of 16 males and 8 females with an average age of 29 years (median: 28.5) and an advanced proficiency with English. Students were paid a sum of £15 for their participation in the experiment. Before introduced to the experimental tasks and systems, the participants were asked to fill out an entry questionnaire so that we could ascertain their proficiency in dealing with multimedia. It transpired that participants have a rather high experience in multimedia searching. The majority of participants deal with image data regularly (once or twice

a day), take photographs occasionally (once or twice a week), and are quite familiar with carrying out image searches.

Most participants stated that their search activities are carried out online, with Google or Yahoo being cited as the most commonly used online services. The photo sharing portal Flickr was named often as well. They mentioned that using these text query based services was generally considered to be easy and satisfactory. One hence noticed different interaction behaviours for different kind of images. Whereas the participants preferred to browse their own images, they feel confident searching for web or other people's pictures by using search queries. They stated that they rarely use photo management tools to organise their personal image collection. The most common practice amongst the participants is creating directories and files on their own personal computer. When asked for the features of an ideal photo management system, the participants stated that the most desired feature was to sort pictures by the date or location they were taken. Another desired feature was to automatically analyse and extend contextual information, where and when images were taken such as surrounding events. Moreover, they would like to have a feature that retrieves images based on a similar visual appearance.

4.3 Results and Analysis

4.3.1 User Perception

On completion of each task provided, participants were asked to describe various aspects of their experience of using each system in post-search questionnaires, by rating the performance of the system on a set of 21 five-point semantic differentials¹ [Heise, 1970]. Four of these differentials focused on the task they had just performed; four focused on the search they had just carried out; three focused on their feeling in interaction with the system during the search; four focused on the set of images retrieved; and six focused on the system itself (see Table 4.2).

In this evaluation, we were interested in feedback on the user satisfaction with the system's features and responses, and the quality of images retrieved from search,

¹ A type of a rating scale between two bipolar adjectives such as good–bad, warm–cold, and bright–dark. It is designed to measure the connotative meaning of objects, events, and concepts, expressing individual's attitudes.

4.3 Results and Analysis

Table 4.2: 21 semantic differentials in post-search questionnaires

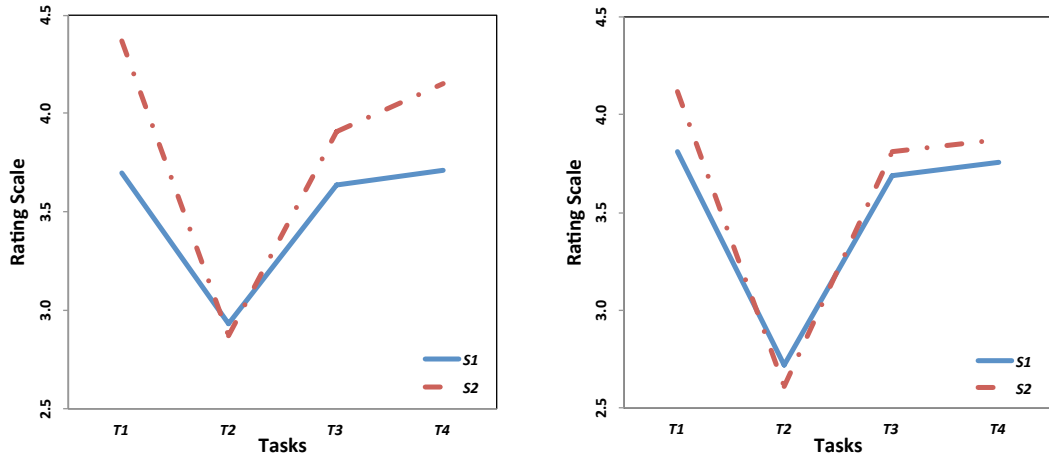
<i>The task were...?</i>			
clear	unclear		
easy	difficult		
simple	complex		
familiar	unfamiliar		
<i>The search was...?</i>			
relaxing	stressful		
interesting	boring		
restful	tiring		
easy	difficult		
<i>While using a system, you felt...?</i>			
in control	not in control		
comfortable	uncomfortable		
confident	unconfident		

<i>The retrieved image set was...?</i>	
relevant	not relevant
appropriate	inappropriate
complete	incomplete
expected	surprising

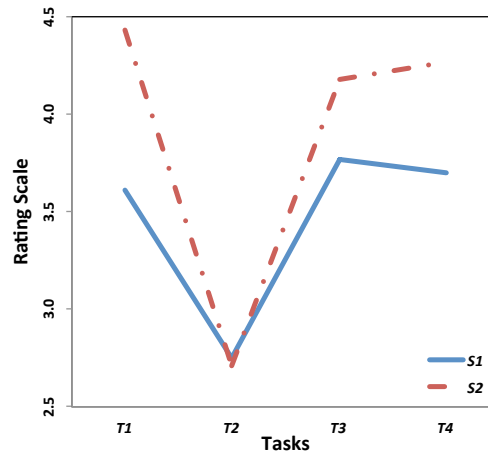
<i>The system was...?</i>	
wonderful	terrible
satisfying	frustrating
easy	difficult
effective	ineffective
flexible	rigid
reliable	unreliable

browsing, and recommendations. For the semantic differentials related to the task performed, the participants stated that the tasks provided were clear, roughly simple, and familiar. However, having analysed the questionnaires by one-way ANOVA, we found that there are significant differences ($p < 0.05$) between the level of task difficulty. It disclosed that task *T2* was the most difficult task followed by tasks *T3* and *T4* whereas task *T1* was the easiest. For other differentials related to our systems, we found that the participants rated the recommender system *S2* slightly better than the baseline *S1* despite no significant difference since the participants felt that they both were effective for solving the task, as they helped them explore the collection, find relevant images, and focus their search. Responses also indicated that the selected images matched what they had in mind before starting the search task and that browsing through the collection made it easy to find relevant images. They stated, however, that the idea of the type of images they were searching for changed whilst performing the tasks. Comments were: “I almost never changed my query word and yet reached many different pictures. So I think the systems works well.”, “I found browsing and recommendations quite efficient, as new aspects or ideas came up in terms of different images.” or “I preferred browsing a lot rather than text searching since browsing helped me in finding more images without posing new queries.”

Aiming to determine the general usability of the system from users’ perspectives, we further asked the participants to judge various statements on a Five-Point-Likert scale from 1 (Disagree) to 5 (Agree). The order of the agreements varies over the



(a) The system was effective for solving the task. (b) The system helped me explore the image collection.



(c) The system helped me discover and define various aspects of the task.

Figure 4.1: Users' satisfaction after interacting with two systems (higher is better), as asked in post-search questionnaires.

questionnaire to reduce bias. We asked them to judge, for examples, the following statements: 1) "The system was effective for solving the task", 2) "The system helped me explore the image collection", and 3) "The system helped me discover and define various aspects of the task". These statements aimed to compare the users' satisfaction of using two different systems to carry out work tasks. Figures 4.1 shows the average judgements of all users on performing four work tasks provided. Considering that all the users interacted with two systems (two tasks on each system), we assume that

they generalised their judgements with respect to the whole system they used rather than the specific features, e.g. text search, browsing, recommendations. Figure 4.1(a) shows the users' agreement that the system effectively support users to fulfil the tasks. Neglecting a drop at the task *T2*, a clear trend towards positive perception for the recommender system *S2* can be observed. The low rating seen in task *T2* can be caused by its difficulty which will be explained in the following section. The similar trend can be observed in Figure 4.1(b) and 4.1(c), depicting the users' opinion about the systems' usability to explore the image collection and various aspects related to the search topics. As can be seen, the users provided a considerably better assessment of the recommender system *S2* than that of the baseline system *S1*. This suggested an overall better performance of the recommendation functionality (as an additional different feature between two systems).

In addition, the post-search questionnaire of the recommender system contained additional questions, asking users about the effectiveness of a particular feature (i.e. the diverse recommendations provided). The averaged answers indicated that they found the recommendations very useful, since the recommendations presented them with images associated with different aspects to solve search tasks. They stated that the recommendations returned a variety of related images that they had not been aware of. Besides, they asserted that the recommendations unveiled some more new aspects of search topics and gave new ideas about how to formulate search queries. They also stated that they often use recommended results to create new browsing aspects. The recommendations helped them find more relevant images with less effort in searching and navigating through the collection. Some quotations: "Recommendations [...] were quite related to images I searched for", "it revealed images that otherwise would not appear" and "the recommendations were easy to manage, they appeared automatically". Few participants, however, said that "sometimes recommendations drew my attention from browsing".

At the end of the user study, we asked all 24 users to evaluate both systems based on various questions. Table 4.3 shows the users' preferences for each of the questions. *S1* denotes the baseline system and *S2* stands for the recommender system. The last column represents a neutral perception about the systems of the users. Whereas only 13.2% of all participants prefer the baseline system *S1*, nearly 50% selected the recommender system *S2* as the best performing system, since it was considered being more

Table 4.3: Users' perception of comparing two systems in an exit questionnaire.

<i>Which system...</i>	<i>S1</i>	<i>S2</i>	<i>=</i>
did you find best overall?	3	13	8
did you find easier to learn to use?	3	9	12
did you find easier to use?	6	7	11
did you prefer?	3	15	6
changed your perception of the task?	1	14	9
did you find more effective?	3	13	8
<i>Percentage</i>	13.2%	49.3%	37.5%

effective and supportive to find new aspects of the task. Even though it provided an additional feature, the participants did not find it more difficult to use the system.

Our analysis of the questionnaires suggests that the participants had more positive perceptions on the recommender system *S2*, which indicates the usefulness of diverse recommendations based on implicit feedback. In a next step, we analysed the resulting log files of their interactions with the interfaces in order to compare the performance of the two interfaces.

4.3.2 Usage Log File Analysis

Agichtein et al. [2006] argued that analysing the users' behaviour whilst using the system can be a valuable source for improving retrieval results. Hence, we assume that the users' behaviour patterns, captured in the log files, can be a strong indicator of the effectiveness of the two image retrieval systems. Assuming that behaviour patterns are directly influenced by the features provided in the graphical interfaces, we expect to identify different patterns for our two interfaces. In the baseline system, users enter search queries and need to perform similar actions on retrieved relevant and non-relevant results; Users will click on the result, browse through the image collection, define new browsing aspects from obtained images, and/or mark the relevant results. The recommender system, however, automatically updates recommendations. Assuming that these recommendations are relevant to the users' information need, they will adopt their interaction strategy accordingly, resulting in a different behaviour pattern with respect to the results.

Table 4.4 shows a mean, standard deviation, and percentage increment of four other measures of exploration, illustrating the user's interaction with the baseline system (*S1*)

4.3 Results and Analysis

Table 4.4: User interaction statistics – mean, standard deviation (in bracket), and percentage increment of *S2* over *S1* (below). No statistical significance at 0.05 level has been found between two systems.

Task	# text queries		# browses		# aspects		# relevant images	
	<i>S1</i>	<i>S2</i>	<i>S1</i>	<i>S2</i>	<i>S1</i>	<i>S2</i>	<i>S1</i>	<i>S2</i>
<i>T1</i>	16.2 (4.4)	12.6 (9.8) -22.22%	11.4 (4.3)	19.1 (8.3) +67.54%	14.1 (3.4)	15.2 (2.8) +7.80%	19.0 (2.3)	19.6 (2.1) +3.16%
<i>T2</i>	22.4 (7.6)	26.3 (9.2) +17.41%	11.5 (2.1)	10.6 (4.3) -7.83%	18.5 (4.2)	17.7 (5.4) -4.32%	13.5 (4.8)	10.9 (6.9) -19.25%
<i>T3</i>	15.8 (4.6)	15.0 (12.4) -5.06%	22.3 (8.0)	23.9 (5.1) +7.17%	13.8 (3.1)	15.3 (4.6) +10.87%	11.9 (7.5)	13.9 (8.9) +16.81%
<i>T4</i>	10.9 (8.0)	9.4 (5.4) -13.76%	14.3 (6.7)	19.7 (6.4) +37.76%	12.8 (3.9)	14.1 (5.1) +10.16%	18.8 (8.4)	19.5 (6.6) +3.72%
Avg	16.3 (6.1)	15.8 (9.4) -3.07%	14.9 (5.8)	18.3 (4.8) +22.99%	14.8 (3.7)	15.6 (4.5) +5.23%	15.8 (11.5)	16.0 (6.1) +1.11%

and the recommender system (*S2*) over all four tasks *T1* – *T4*. The first column denoted “# text queries” shows the number of unique text queries executed on search. The second column denoted “# browses” lists the number of clicks user performed for browsing using the different interfaces. The next column denoted “# aspects” shows the number of aspects (tabs) created for different exposed aspects. The last column denoted “# relevant images” depicts the total number of relevant images added to the presentation panel.

An one-way ANOVA analysis of the results did not reveal any significant differences between the number of browses, queries, sessions, or results for the two systems. Nevertheless, the results suggested that diverse recommendations can improve an effectiveness of image browsing systems. As Table 4.4 shows, participants clicked on images for browsing (on average) in the recommender system *S2* more than in the baseline system *S1*. Vice versa, the users entered fewer text queries in the system *S2* than in the system *S1*. These results suggested that the system *S2* assists the users to rely more on browsing and less on text search functionalities. In three out of four tasks, i.e. *T1*, *T3*, and *T4*, the users put less effort in formulating search queries whereas finding more relevant images.

Although the results of task *T2* contrast with those of the other tasks, this might be due to the level of task difficulty. The questionnaires and log analysis are rather concordant. Task *T2* was perceived as the most difficult task, followed by tasks *T3*, *T4*, and *T1*. In task *T2* the users performed the higher number of text queries, clicked less

for browsing, and found the fewer number of relevant images. This suggested that the browsing functionality was not suitable for the task *T2* and therefore recommendations, which depends on implicit feedback extracted from browsing interactions. One of the main problems in task *T2* was that the users found it difficult to formulate an initial search query, which would retrieve useful results to be used for beginning browsing (i.e. query by image example). Users spent most time finding example images from text search, and used them to browse through a collection. As a result, recommendations, which are based on browsing interactions, were used less and could not provide many results to the users. This is a drawback of this browsing technique that requires example images to start browsing. We do not however focus our research on this issue as we want to investigate the benefit of providing diverse results in exploratory search tasks.

Another possible reason for the results in the tasks *T2* was the level of specification for given topics due to the nature of the collection. Task *T2* might be the “narrowest” in comparison to tasks *T1*, *T3*, and *T4*. To explain this, we analysed users’ agreement between the set of relevant images, assuming that there will be less agreement amongst users for broader tasks, which require a greater extent of interpretation. For task *T2*, 38.6% of unique results were selected by two or more users. For tasks *T1*, *T3*, and *T4*, two or more users selected 29.0%, 29.9%, and 27.6% respectively. The greater number of agreement amongst users in task *T2* is consistent with task *T2* being the most specific task.

Aiming to evaluate the effectiveness of *diverse* recommendations in supporting information exploration, we analysed the number of defined aspects in given search tasks. As Table 4.4 indicates, the participants created more aspects (on average) in the recommender system *S2* than in the baseline system *T1*. Only in the task *T2*, the number of aspects defined by the users of the system *S2* is little fewer (-4.32%) than the system *S1*. This might be again (as discussed above) the nature of the task *T2* that differs from the other three tasks. From these results, it suggested that the diverse recommendations in the system *S2* help the users discover more new aspects related to search tasks.

In Table 4.5, we show the average number of images used to define aspects (top) and marked as relevant (bottom) in the recommender system *S2*. These images were

Table 4.5: No. of images (in percentage), obtained from text search, browsing and in particular recommendations, exploited to define aspects and marked as relevant in a recommender system (S2).

	task	# avg. images exploited (100%)	text search	browsing	recommendations			
					total	CLD	EHD	HTD
aspects	T1	15.2	10.7%	71.2%	18.1%	6.8%	9.3%	6.2%
	T2	17.7	43.1%	42.3%	14.6%	3.7%	5.7%	13.3%
	T3	15.3	12.3%	69.1%	18.6%	7.1%	6.6%	8.7%
	T4	14.1	15.0%	69.3%	15.7%	7.6%	7.6%	5.9%
	Avg	15.6	20.3%	63.0%	16.8%	6.3%	7.3%	8.5%
relevant images	T1	19.6	7.1%	59.7%	33.2%	15.8%	12.8%	14.0%
	T2	10.9	70.3%	23.0%	6.7%	1.6%	3.3%	2.5%
	T3	13.9	10.3%	46.0%	43.7%	16.2%	17.4%	19.8%
	T4	19.5	13.3%	63.3%	23.4%	8.6%	10.8%	7.7%
	Avg	16.0	25.2%	48.0%	26.8%	10.6%	11.1%	11.0%

the results from text search, browsing, or recommendations. We also reported the results in percentage of which respective search methods they come from. Further, the table shows which low-level feature was used to produce recommendations exploited by users. The abbreviations stand for colour layout descriptor (CLD), edge histogram descriptor (EHD) and homogeneous texture descriptor (HTD). Recommended images can accrue from the union of different low-level features. The total number of recommendations in the table is hence smaller than the sum of the presented features.

As can be seen, roughly every sixth defined aspect was based on recommendations. It also shows that almost one-quarter of all images that were marked as relevant came from the recommendation panel. This suggested that the participants often relied on the provided recommendations. Furthermore, comparing these proportions with the percentage increment of the used images in the Table 4.4, recommendation was the main functionality that increases the performance of the system S2 over S1. This is because, for instance, on average $15.6 \times 16.8 / 100 = 2.62$ recommended images were used to define aspects whereas users defined $15.6 - 14.8 = 0.8$ more aspects when using the system S2. Hence, the improved performance of the systems S2 is likely to be from the recommendation functionality as the former is higher than the latter, suggesting that it does not happen by chance or from other functionalities. Similar results can be observed from the number of relevant images marked by users, where the recommendation based on implicit feedback assists the users to find more relevant results covering many aspects.

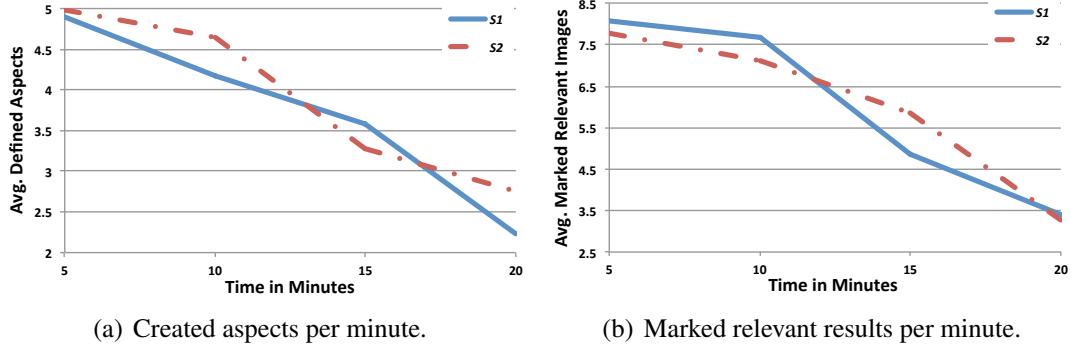


Figure 4.2: Users' interaction patterns over the course of an experimental session when using two systems.

Additionally, Table 4.5 shows that users did not prefer recommendations from any specific low-level features, since they relied equally on recommendations visually diversified by different features, i.e. colour, edge, and texture. This suggests that the diverse recommendations relieved the participants from relying on the results from one low-level feature only. On the contrary, if other features were not helpful for users, they would have used the recommendations obtained from a specific feature only. We therefore conclude that a diversity is a useful means to present recommendations to the users.

Moreover, we were interested in analysing how the participants interacted with both systems of various time points during their search sessions. Figures 4.2(a) and 4.2(b) show the numbers of created aspects and the number of images that were marked as relevant using both systems, respectively. The two figures reveal an interesting search pattern. In the first ten minutes of the search session, the participants created more aspects in the recommender system S2, but at the same time marked fewer relevant images. After 15 minutes, however, this pattern changed towards creating more aspects using the baseline system S1 and marking more relevant images using the recommender system S2. At the end of the search session, the pattern reversed again.

4.4 Findings and Discussion

This chapter has investigated the benefits of diversity in search results from the users' perspective in exploratory search. We analysed both the questionnaires and usage log files of our user study, aiming to answer research questions defined in Section 4.2.2.

Participants were asked to express their opinion about the usefulness of two systems in fulfilling the tasks by filling in interim questionnaires at various stages of the user study. These questionnaires were analysed under four main criteria as follows.

First of all, we looked at the user perception in semantic differentials after using two different systems to perform each tasks. The participants' responses indicate that they found two systems are rather easy, satisfying, and reliable whereas the recommender system were rated slightly more effective than the baseline system to solve the tasks. Next, we evaluated the general usability of the systems and the way in which the systems support the users. In three out of four tasks, participants agreed with the statements in the post-search questionnaires that the recommender systems are more effective and helpful to carry out the tasks, explore an image collection, and discover more and various aspects of a search topic. This basically shows the usefulness of diverse results as provided by the recommender system from the users' perspective. Although in one of four tasks the recommender system seems to be less useful, the baseline also fails when additional helps, i.e. diversified results, are required. This is a limitation of this type of recommendation which relies on implicit feedback from the feature of baseline system, i.e. clicks for browsing (# browses in table 4.4). In particular, when such a feature does not work as expected, the recommendation will not be able to work as well.

Further, the agreements were supported by the open question about the recommendation quality, as asked in the post search questionnaire of the recommender system. Some participants stated, for example, that "In general, the recommendation is great to explore an image collection according to the user interests. It automatically shows various images that interest me over the time I used the system. I did not need to search again using the keyword box. I just clicked on images for browsing and the recommendation gave me the results immediately." Finally, we asked the participants to compare two systems in the exit questionnaire of the experiment. The users' responses suggests that the recommender system is preferred to the baseline system, as almost 50% of users voted for the recommendation, in comparison with 13.2% for the baseline and 37.5% for a neutral perception. We derived these findings from questionnaires to answer the first research question **RQ1**.

Aiming to answer the other research questions **RQ2-4**, we analysed log files of user interactions with the systems. As we have shown in the previous section, users discov-

ered and defined more aspects when using the recommender system. This suggested that diverse recommendations are helpful for users to deal with the tasks associated with multiple aspects. We also found that users gathered more relevant images in the recommender system. The further analysis also suggested that the percentage increments of aspects and relevant images that users found mainly came from the recommendations. We therefore conclude that diverse recommendations are useful to users when they have a multi-aspect information need. Moreover, as a higher number of relevant images were found in the recommender systems, we conclude that implicit feedback technique can successfully be employed to provide relevant recommendations.

From the results of our study, we derived the following findings which answer research questions **RQ1-4**.

- 1) Considering a set of questionnaires, user's responses indicate that diverse recommendations are useful for users to complete complex work tasks, associated with multiple aspects.
- 2) By analysing usage log files, we see that users discovered more aspects of a work task when using the recommender system.
- 3) Users mainly defined aspects from browsing results. However, recommendation is the main feature that increases the performance of the recommender system in helping users explore more aspects.
- 4) As can be observed, users found more relevant images in the recommender system and thus implicit relevance feedback is an effective source to generate the recommendation relevant to user information need.

4.5 Summary

In this chapter, we aimed to confirm the benefit of result diversity from the users' perspective. We investigated this issue from the evaluation of diverse recommendations that have been outlined in Chapter 3. We employed a user-centred evaluation where 24 participants were asked to carry out complex work tasks using the Ostensive Browser Plus system that contains the recommendation functionality. The recommender system

exploits implicit relevance feedback mined from user browsing trails to generate a set of potentially relevant images to recommend. A hierarchical clustering technique is then applied on different visual features to select and diversify recommendation. The recommendation allows users to explore a data collection to a greater extent, presenting documents covering various aspects of image contents.

We evaluated four research questions by analysing user preferences and interactions which were provided during various stages of the experiment. The analysis revealed that diverse recommendations are effective to support users in their exploratory search tasks. Users discovered and defined more aspects in our aspectual browsing interfaces when using the recommender system. Further, the recommendation assists users to gather more relevant images for the tasks. Regarding the users' response given in questionnaires, the recommender system is preferred to the baseline system and diverse recommendations are a welcome feature that users prefer to be included in IR systems, in particular when they have to deal with the work tasks related to multiple subtasks or aspects. We therefore conclude that diverse recommendations are in effect beneficial to users, and implicit relevance feedback can be used to capture users' interest and to recommend relevant image documents. These findings suggest that result diversity brings substantial benefits to users when they have a multi-aspect information need.

Part III

Ranking Paradigms for Result Diversification

Chapter 5

Ranking Paradigms and their Integrations for Sub-topic Retrieval

5.1 Introduction

In Part II, we investigated the advantages of providing result diversity from the users' point of view. It was found that users benefit from the diversity when they have a multi-aspect information need. We proposed an interactive retrieval system that provides diverse recommendations and addressed the problem of aspect retrieval¹. The purpose of the system is to support users in finding documents, which cover different *aspects* of their information needs.

Here, we study a particular class of *automatic* methods for a retrieval problem that is intimately related to that of aspect retrieval. In particular, we consider the problem of sub-topic retrieval, developing methods for producing a ranked list of documents that provide a complete coverage of sub-topics². To this aim, we study the sub-topic retrieval problem by retaining the basic “query-in, ranking-out” model traditionally employed in IR. We seek the methods that modify the ranking in order to include as many relevant sub-topics as possible at early ranks. Once the suitable methods become

¹The problem of aspect retrieval is investigated in the TREC interactive track. For the users of interactive retrieval systems, the objective is to find as many relevant documents as possible, so that taken together they cover as many different *aspects* of the topic as possible [Over, 2001; Voorhees and Harman, 2001].

²Sub-topic is a common term used in the context of sub-topic retrieval, where the problem has to deal with the ambiguity of queries and the uncertainty about users queries which can refer to many aspects. The goal of this context is to model *dependent relevance* and find documents that cover as many sub-topics of a general topic as possible [Zhai et al., 2003].

available, then they can be employed within interactive retrieval systems, where sub-topical diversity is of major concern and regardless of the type of result presentation such as a ranked list of retrieval results or a set of recommendations.

The notion of relevance is central to information retrieval models. In IR, the relevance of a document is typically assumed to be independent of the relevance of other documents. This assumption is on the basis of the Probability Ranking Principle (PRP) [Robertson, 1977], which enables retrieval systems to estimate the relevance of each document separately. Documents are ranked exclusively according to their probability of being relevant to a query. By adhering to such retrieval policy, it is likely that the list of documents retrieved by PRP addresses only a particular aspect of the information need [Stirling, 1981]. In real search scenarios, however, the independent relevance assumption often does not hold and consequently ranking approaches that rely on it, such as the PRP, provide *sub-optimal* document rankings regarding the expected utility [Gordon and Lenk, 1992].

To overcome the limitations of the independent relevance assumption, some efforts have been devoted to the development of dependent relevance models. In parallel, other approaches have been devised to predict sub-topics, estimated by the relationship between documents. These approaches can be thought of as two faces of the same coin: generally, diversifying a document ranking implies exploiting document dependencies; and, vice versa when accounting for document dependencies to model sub-topics (at relevance level), diversification can be achieved.

Two different patterns can be recognised from the approaches suggested in the literatures for the purpose of ranking diversification, i.e.:

- 1) Inter-dependent document relevance paradigm.
- 2) Sub-topic aware paradigm.

5.1.1 Inter-dependent Document Relevance Paradigm

When ranking documents, relationships between documents are taken into account by promoting documents that differ from each other. These approaches maximise, at each rank position, a function that depends upon both relevance estimates and documents relationships. The intuition underlying this is that diversity can be achieved by ranking

relevant documents based upon the novelty of their contained information. A similarity function is usually employed to estimate the novelty of a document (the less a document is similar to those already ranked, the more it carries novel information). Examples of heuristic or theoretically driven approaches that include document dependencies in the ranking function are:

- *maximal marginal relevance (MMR)* [Carbonell and Goldstein, 1998], which interpolates document relevance and documents relationships;
- *modern portfolio theory (MPT)* [Wang and Zhu, 2009], which combines relevance estimates and document correlations;
- *quantum probability ranking principle (qPRP)* [Zucon et al., 2009a], which implicitly captures dependencies amongst documents through quantum interference; and
- *interactive probability ranking principle (iPRP)* [Fuhr, 2008; Zucon et al., 2011a], which includes document dependencies through user interactions.

Without incorporating document dependencies in the ranking function, an inverse approach that can be classified into this category is based on the *pruning technique*, where documents that are too similar to other documents (greater than their threshold θ) are *removed* from a result list [Carterette and Chandar, 2009].

5.1.2 Sub-topic Aware Paradigm

The need of (sub-topical) diversity can be satisfied by directly modelling sub-topics from documents, with the assumption that they are *all* relevant and correspond to the user information needs. Regardless of document relevance, relationships between documents are employed to estimate sub-topics. The rationale behind this paradigm is that closely associated documents tend to fulfil similar information needs and thus retrieving documents that belong to different groups of similar documents is likely to satisfy different information needs. Example techniques that attempt to predict sub-topics from documents are:

- *clustering* [MacQueen, 1967],

- *classification* [[Huang et al., 1998](#)],
- *latent Dirichlet allocation (LDA)* [[Blei et al., 2003](#)],
- *probabilistic latent semantic analysis (PLSA)* [[Hofmann, 1999](#)], and
- *relevance models* [[Carterette and Chandar, 2009](#)],

whereas other related techniques using external resources, e.g. Open Directory Project taxonomy¹ or query log, to model sub-topic are:

- *intent aware select (IA-select)* [[Agrawal et al., 2009](#)], and
- *query features* [[Santos et al., 2010](#)].

Afterwards, documents are diversified by, for instance, interleaving in a ranking list the documents that belong to different estimated sub-topics, or interpolating document relevance and sub-topic estimates. Several ranking criteria can be applied to select documents after the information about the estimated sub-topics has been obtained.

5.1.3 Goal and Plan of the Chapter

In this chapter, we aim to understand how the current approaches based on those two paradigms actually perform on retrieval systems to promote diversity and how they compare with each other for the same purpose of result diversification. In order to develop a better retrieval strategy, it is instructive to understand how different approaches achieve the same goals of sub-topic retrieval, i.e. providing a complete array of sub-topics and meanwhile avoiding excessive redundancy in search results. To this aim, this chapter is devoted to review a number of state-of-the-art approaches for diversifying documents. We outline representative approaches, such as MMR, MPT, clustering, LDA, and PLSA, selected from the two categories. These approaches will then be empirically investigated in Chapter 6. Furthermore, we focus just on the diversification methods driven by the intrinsic characteristics of the search results, i.e. documents' content or coverage of particular sub-topics. These methods are opposed to those, e.g.

¹www.dmoz.org

suggested by Santos et al. [2010], which use external information to predict intents of user queries, e.g. query log, ontology, Wikipedia¹ and DBPedia².

Moreover, we investigate whether a new ranking approach can be devised so that we can integrate the benefits of the two ranking paradigms into a unified framework, regardless of the choices of similarity estimation function, document dependency function, and sub-topic modelling algorithm. We propose a general result diversification framework based on the *integration* approach, which explicitly models clusters of documents with respect to sub-topics and ranks documents by including dependencies between documents in a result list. Our framework enables the development of a variety of algorithms for integrating statistical similarity and diversity structures, conveyed by sub-topic clusters and document dependencies [Leelanupab et al., 2010a,b,d].

The rest of this chapter is organised as follows. The next section briefly describes the PRP and its limitations. Then, we review existing retrieval approaches for encoding novelty and diversity in document ranking. In Section 5.4, we illustrate our framework based on inter-document dependencies and sub-topic evidences induced from document clusters in order to provide better search results for sub-topic retrieval task. Finally, we summarise this chapter in Section 5.5.

5.2 The Probability Ranking Principle

Maron and Kuhns [1960] suggested that documents should be ranked in the order of the probability of relevance to the request, or of usefulness to the user, or of satisfying the user. The concept of probability of relevance was introduced due to the fact that no retrieval system can be expected with *certainty*, which documents a user might find useful, i.e. only the user can judge the relevance or usefulness of documents. Therefore, the system must necessarily deal with the *probability* and the information retrieval system should be designed accordingly. The estimation of such probability has become a major area of IR research. Several approaches have been proposed to estimate this probability, such as the 2-Poisson model [Bookstein and Swanson, 1974], the Binary Independence model [Robertson and Spärck-Jones, 1976], and the BM25 model [Robertson et al., 1994].

¹<http://en.wikipedia.org>

²<http://dbpedia.org>

5.2 The Probability Ranking Principle

The *probability ranking principle (PRP)* is the most well-accepted ranking theory in information retrieval and is commonly attributed to Robertson and Spärck-Jones [1976]. This principle states that documents should be ranked and presented to a user in descending order of document's probability of relevance. The PRP has been proven optimal from a theoretical point of view [Robertson, 1977] and can be justified using utility theory [Gordon and Lenk, 1991]. It yields the maximum expected number of relevant documents, and thus maximises the values of the set-based measures such as precision and recall. Moreover, the optimal performance can also be expressed in terms of costs associated with the retrieval of non-relevant documents and the non-retrieval of relevant documents. The definition of PRP makes a number of assumptions. Specifically:

- I) Relevance (or usefulness, or user satisfaction) is a dichotomous judgement, i.e. a document can only be judged either relevant or not-relevant, and there are no in-between decisions.
- II) PRP is applied only to a single request and not to a set of requests, i.e. a series of reformulated queries issued by a user during search sessions.
- III) The *relevance* of a document to the user is independent of the other documents in the corpus (we refer to this as *independence assumption*).

In particular, assumption 3 is the key concept of PRP, which forms the theoretical basis for probabilistic retrieval models. It assumes that a user's relevance assessment of a document will not change during the course of information seeking. It implies that systems expect a user to judge the relevance of a document in isolation with other documents that he/she has already seen. This is also the case for the relevance judgements commonly made in TREC, where documents are assumed to be independently judged [Voorhees and Harman, 2001]. As a result, the relationship between query and document become both necessary and sufficient to establish relevance [Goffman, 1968].

Formally, given a query q , if $P(R|x_i, q)$ is the probability of relevance estimated for document x_i , then the PRP suggests to present at rank $J + 1$ a document x such that:

$$PRP_{J+1} \equiv \operatorname{argmax}_{x_i \in I \setminus J} [P(R|q, x_i)] \quad (5.1)$$

where I is the set of results retrieved by the IR system; J is the set formed by the documents ranked until iteration J ; x_i is a candidate document in $I \setminus J$, which is the set of documents that have not been ranked yet.

Nevertheless, the independence assumption has been brought into question by [Cooper \[1976\]](#); [Goffman \[1964\]](#). They showed counter-examples to the PRP, e.g. the situation in which the relevance or usefulness of one document affects the relevance or usefulness of another. Suppose a user does *not* assess documents individually, and presumably examines documents in a sequential order of document presentation in a ranking, such as x_1, x_2, \dots, x_N . If document x_1 is judged relevant by the user, it may provide some indication of the possible relevance of x_2 . That is, the relevance of document x_1 may affect the document x_2 , provided that x_2 simply repeats information contained in x_1 . As a result, document x_2 may be judged not-relevant by the user if examined after x_1 . This example, however, is not valid when documents x_1 and x_2 are neither relevant on their own, but they have to be relevant together since each provides complementary aspects of the problem. Regardless of the above case, the relevance of a set of documents does not only depend on the individual relevance of itself, but also depends on the relationship between the documents. In particular, this counter-example has been supported by the study of [Eisenberg and Berry \[2007\]](#), where relevance scores assigned to documents are influenced by the order of document presentations in a common relevance assessment activity.

A number of empirical studies have suggested that PRP cannot be extended to all retrieval scenarios in information retrieval [[Agichtein et al., 2006](#); [Boyce, 1982](#); [Stirling, 1977](#); [Zhai and Lafferty, 2006](#)]. An example scenario was investigated by [Chen and Karger \[2006\]](#), where the user is satisfied with a few number of relevant documents rather than trying to find all relevant documents. They showed that in such a scenario, it is more effective for a retrieval system to optimise document ranking in a way that the probability of finding at least one relevant document amongst the top n is maximised. They also found that by doing so, the diversity of documents amongst the top n documents is inherently promoted in terms of *instance recall*¹.

¹The original measure used in the interactive track of TREC [[Goffman, 1964](#)] inspires the development of sub-topic recall [[Zhai et al., 2003](#)].

In addition, [Gordon and Lenk \[1991, 1992\]](#) showed that a traditional ranking criterion of PRP provides a sub-optimal ranking when specific measures are used to define the optimality of document ranking. This is because the PRP ignores the uncertainty when estimating the probabilities of documents' relevance and the correlation between such probabilities. This is especially true in the case of *sub-topic retrieval*, where there is the need of accounting for document dependencies, and thus ultimately for diversity when ranking document is of importance. The relevance encoded by query-document relationship is insufficient to determine the “utility” of a document. Instead, the relationships between documents should be included to measure the utility. In this case, PRP does not provide a satisfiable ranking because it discards the dependencies between assessments of document relevance. This is known as the limitation of the PRP, and, although it does not affect the optimality of the ranking principle for tasks such as ad-hoc retrieval, it is the cause for the sub-optimality of the PRP particularly in sub-topic retrieval.

5.3 Background of Result Diversification

5.3.1 Beyond Independent Relevance

The independence assumption of PRP has been recognised as an important issue in information retrieval. [Zhai et al. \[2003\]](#) argued that it is *insufficient* to return a set of relevant documents where the relevance of a document is treated independently from that of other retrieved documents. This is because the utility of retrieving one document, in general, may depend on which documents a user has already seen. An extreme example is the situation where a relevant document may become useless if the user has already seen documents with the same content. Their observation gives rise to new evaluation measures and retrieval strategies that consider dependencies amongst documents.

Many recent works attempt to overcome PRP's limitations by including document dependency in the ranking function [[Carbonell and Goldstein, 1998](#); [Fuhr, 2008](#); [Wang and Zhu, 2009](#); [Zuccon et al., 2009a](#)]. These approaches share a common structure as they include not only *relevance* estimations but also *diversity* estimations. The diversity estimations are used to measure the degree to which pairs of documents differ. To

obtain the final document relevance, or what we call *inter-dependent document relevance*, the two estimations are combined by a composition function (e.g. addition, multiplication, etc.). In this section, we examine two popular examples of ranking approaches for sub-topic retrieval based on the inter-dependent document relevance paradigm: MMR and MPT.

5.3.1.1 Maximal Marginal Relevance

A simple and intuitive method to address diversity between documents is that of *maximum marginal relevance (MMR)* [Carbonell and Goldstein, 1998]. Using a tuneable parameter, this ranking method balances the relevance between a candidate document and a query, e.g. the probability of relevance, and the similarity between the candidate document and all the documents ranked at previous positions. The ranking is linearly produced by maximising relevance and inter-document similarity at each rank. The MMR strategy is characterised by the following ranking function:

$$MMR_{J+1} \equiv \operatorname{argmax}_{x_i \in I \setminus J} [\lambda S(x_i, q) - (1 - \lambda) \max_{x_j \in J} S(x_i, x_j)] \quad (5.2)$$

where I is the set of documents retrieved by the traditional ranking method, e.g. BM25 or language model; J is the set of documents that have already been ranked, i.e. x_j ; and x_i are candidate documents in $I \setminus J$, which is the set of documents that have not been ranked yet. The function $S(x_i, q)$ is a normalised similarity metric estimating the relevance of document x_i to a query q , and $S(x_i, x_j)$ is a similarity metric estimating the redundancy between a pair of documents, i.e. documents x_i and x_j . In other words, $S(x_i, x_j)$ is used as an indicator of *novelty* (i.e. the fewer pairs of the candidate document x_i against all other documents x_j are similar, the more novel information the document x_i contains with respect to others). λ is a hyper-parameter that linearly combines $S(x_i, q)$ and $D(x_i, x_j)$. The hyper-parameter λ can be inferred by the user's model, which characterises users with preference for documents rankings conveying the amount of relevant or novel information. A value of λ greater than 0.5 assigns more importance to relevance than to novelty/diversity. Conversely, when $\lambda < 0.5$, novelty is favoured over relevance. In other words, if λ is equal to one, the ranking criterion is

the same as those of PRP, rejecting any evidence provided by the novelty estimation, or only relevance is considered.

In our experimental study, when operationalising MMR, we modify how the novelty function, $\max_{x_j \in J} S(x_i, x_j)$ in equation (5.2), impacts on the ranking. We substitute the function *max* with *avg*, which returns the average similarity value between all pairs of x_i and x_j , instead of their largest value. By doing so, the similarity values of all pairs are considered in the ranking function rather than a single pair that has the highest similarity value. The underlying intuition is that we want to retrieve the relevant documents with the contents which are different from all the documents that have already been ranked. To compute the novelty contained in a candidate document with respect to other previously ranked documents, we here use the cosine similarity metric to measure the similarity between documents' term vectors obtained by the BM25 weighting scheme. We can estimate the similarity between a pair of document vectors using the following formula:

$$\text{avg}_{x_j \in J} S(x_i, x_j) = \frac{1}{|J|} \sum_{j=1}^{|J|} (S(x_i, x_j)) \quad (5.3)$$

In Figure 5.1(a) we depict the document selection procedure suggested by MMR. Note that this method does not actually perform clustering, but we simulate and outline *imaginary* clusters in the figure (as highlighted by circles in dash lines). The imaginary clusters identify the possible sub-topics covered by those documents. As shown in the figure, documents inserted in the ranking following MMR might belong to the same cluster (i.e. x_1 and x_3 from a gray cluster), contrary to what is required in the sub-topic retrieval task, i.e. return documents that cover as many sub-topics as possible.

5.3.1.2 Modern Portfolio Theory

Wang and Zhu [2009] suggested to rank documents according to a paradigm proposed in the theory of financial investment, the *modern portfolio theory (MPT)*, which maximises the returns on expected investment portfolio for an acceptable level of risk. In the IR scenario, diversification is achieved using MPT by reducing the risk associated with document ranking. The intuition underlying MPT is that an ideal ranking order is

the one that balances the relevance of a document against the level of its risk or uncertainty (i.e. variance). Thus, when ranking documents, relevance should be maximised whilst minimising variance. The objective function that MPT optimises is:

$$MPT_{J+1} \equiv \operatorname{argmax}_{x_i \in I \setminus J} \left(P(R|q, x_i) - bw_{x_i}\sigma_{x_i}^2 - 2b \sum_{x_j \in J} w_{x_j}\sigma_{x_i}\sigma_{x_j}\rho_{x_i,x_j} \right) \quad (5.4)$$

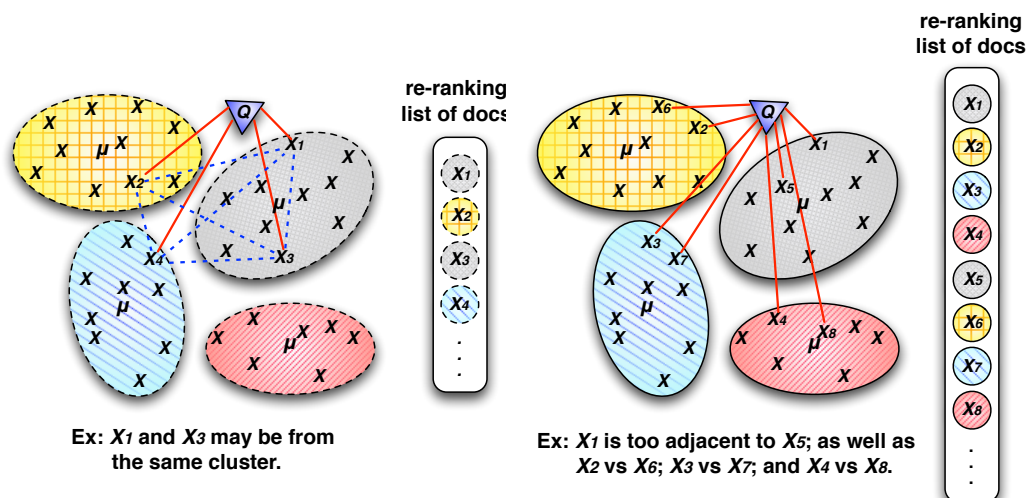
where $\sigma_{x_i}^2$ is the variance associated to the probability estimation of document x_i , and ρ_{x_i,x_j} is the Pearson's correlation between document x_i and document x_j . Besides, b is a parametric coefficient representing the risk propensity of a user and w_{x_i} is a weight, inversely proportional to the rank position, expressing the importance of the rank position. In particular, $b < 0$ represent the situations where users are inclined to accept the risk, whereas $b > 0$ represent the situation where users are risk averse. Finally, if $b = 0$, then only the relevance estimation is considered, resulting in a PRP-like ranking criterion.

In summary, MMR and MPT have a similar underlying additive schema for combining relevance and diversity. A common component of their ranking functions is the estimation of the probabilities of relevance. Both methods then balance the relevance estimation using a second component, which in turns captures the degree of diversity between the candidate document and the ranking. Other approaches that implement, to some extent, the inter-dependent document relevance paradigms have been proposed: see for example the seminal work of Goffman [1968] and Zuccon et al. [2009a]. In the empirical study presented in Chapter 6, we implement both MMR and MPT and compare them with our proposed framework.

5.3.2 Sub-topic Aware Paradigm for Diversity

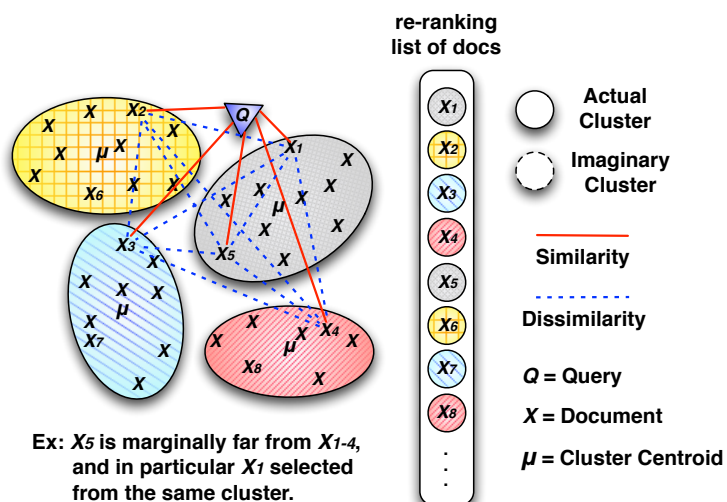
A number of methods belonging to the sub-topics aware paradigm derive from the topic-based approach for information retrieval. This paradigm employs an informative prior knowledge based on the topical content of documents to model a hypothetical set of sub-topics, represented by the clusters of documents. The rationale behind this paradigm is that, by grouping documents with similar (sub)-topical contents into the same cluster, the selection of documents from different clusters should potentially produce diverse results obtained by selecting documents with different contents.

5.3 Background of Result Diversification



(a) MMR with imaginary clusters.

(b) Clustering with document selection by document relevance.



(c) Clustering with document selection by inter-dependent document relevance (e.g. MMR).

Figure 5.1: Re-ranking methods for promoting diversity.

There have been other associated notions derived from cluster-based retrieval models, where (sub)-topic modelling can be viewed as an application of clustering to improve the ranking effectiveness in ad-hoc retrieval. The basis of this approach is the well-known *cluster hypothesis*, which states that “closely associated documents tend to be more relevant to the same requests”¹ [Hearst and Pedersen, 1996; van Rijsbergen, 1979]. In other words, cluster-based retrieval assumes that the probability of relevance of a document depends on the relevance of other *similar* documents to the same query. Therefore, relevant documents are likely to be more similar to each other than to non-relevant documents, and thus they are likely to be clustered together. Furthermore, similar documents tend to fulfil similar information needs, and these may reflect different sub-topics the user might be interested in. Therefore, by selecting documents from different clusters, more relevant documents can be found and more diverse sub-topics can be promoted to the top of the ranked list.

Traditionally, document clustering is applied over the whole collection of documents prior to querying. On the other hand, another interesting type of clustering that is only applied to the search results of a query is *query-specific clustering*, or *query-biased clustering*. Here, clusters are constructed from the top-ranked (e.g. 100) results in response to a given query. While search result clustering has been shown to improve retrieval effectiveness in ad-hoc retrieval [Kurland, 2006; Kurland and Domshlak, 2008; Liu and Croft, 2008; Tombros et al., 2002], we aim to examine the effectiveness of query-specific clustering in sub-topic retrieval. In particular, query-specific clusters are intuitively appealing since they have the potential to represent sub-topics of a topic/query.

Several techniques can be employed to infer which sub-topics are likely to be covered by documents. For example, in [Carterette and Chandar, 2009], latent Dirichlet allocation (LDA) [Blei et al., 2003] and Lavrenko’s relevance models [Lavrenko and Croft, 2001] are employed for estimating the presence of sub-topics within documents. Alternative techniques that may be employed to this end are probabilistic latent semantic analysis (PLSA) [Hofmann, 1999] and clustering (e.g. K-means clustering [MacQueen, 1967]).

¹Note that this hypothesis does not actually mention cluster. However, “closely associated” or similar documents will generally be in the same cluster.

Regardless of the specific techniques employed to estimate sub-topics, document rankings that exploit such explicit evidence induced by clustering can be formulated in various ways. In cluster-based retrieval, [Kurland and Lee \[2004\]](#) introduced an *interpolation* algorithm that enhances retrieval performance of document-based language models by incorporating individual-document and cluster information. In sub-topic retrieval, [Carterette and Chandar \[2009\]](#) proposed faceted retrieval model, where sub-topics are called facets in their work. LDA is used to capture a set of facets, and a subset of documents is selected in such a way that their marginal likelihood is maximised. Regardless of document relevance, the likelihood is computed by the probability that a facet is covered by a document. In IA-select [[Agrawal et al., 2009](#)], the selection of document is determined by its relevance to the query as well as the probability that it satisfies potential query-intents (sub-topics) given that all previously retrieved documents fail to do so. Within the image retrieval domain, [Deselaers et al. \[2009\]](#) and [Zhao and Glotin \[2009\]](#) used a straightforward technique based on round-robin cluster selection. Their technique assumes that presenting documents which belong to different clusters is a means to guarantee the diversity of sub-topics in a ranking. However, none of these approaches include document dependencies when ranking documents in a result list. This issue motivates us to develop our framework which integrates two retrieval paradigms for result diversification.

In summary, approaches belonging to the sub-topic aware paradigm consist of two separate steps:

- 1) Document clustering; and then
- 2) Selection of documents based on the cluster structure.

In the following subsections we outline three sub-topic modelling techniques we experiment with for diversification, i.e. K-means clustering, PLSA, and LDA. Then, we describe two approaches for document selection to exploit information from sub-topics as estimated by clustering similar documents.

5.3.2.1 Sub-topic Modelling Techniques

K-means. A simple method to generate groups of similar documents is that of K-means clustering [[MacQueen, 1967](#)]. When clustering documents, K-means algorithm

aims to iteratively partition the N documents into K clusters ($K < N$) in which each document belongs to the cluster with the nearest mean. Documents are represented by feature vectors of term weights, e.g, *tf.idf*, BM25 weights, etc. The number of clusters K remains constant from the beginning to the end of the algorithm. During each iteration, each document is kept in the same cluster or assigned to a different cluster if the new minimum of all the K distances is obtained. This process is repeated until the stopping criteria is met. That is, finding cluster assignments for documents that achieve a global minimum of the *objective* function Obj . Given a set of documents $X = \{x_1, x_2, \dots, x_N\}$, the objective function is defined as:

$$Obj \equiv \underset{C}{\operatorname{argmin}} \sum_{k=1}^K \sum_{x_i \in c_k} \|x_i - \mu_{c_k}\|^2 \quad (5.5)$$

where $C = \{c_1, c_2, \dots, c_K\}$ is the set of clusters, μ_{c_k} is the geometric centroid of cluster c_k , and $\|x_i - \mu_{c_k}\|^2$ is a chosen distance measure, such as the Euclidean distance between x_i and μ_{c_k} .

PLSA and LDA. Statistical methods that provide topic modelling, e.g. probabilistic latent semantic analysis (PLSA) [Hofmann, 1999] and latent Dirichlet allocation (LDA) [Blei et al., 2003], can be used to discover the abstract “topics” that occur in a collection of documents, and in our case to discover the “sub-topics” in search results. These methods are the generative probabilistic models for discrete collection used for textual data.

Within the probabilistic framework, PLSA and LDA represent documents as a distribution probability over latent topics, where each latent topic (that in the context of our study provide the evidence for forming clusters) is a distribution over words. Once the latent topic models have been generated, it is possible to infer the probability of topics contained in a document based on arbitrary words within it. To perform clustering with PLSA and LDA, the topic models are trained over a set of retrieved documents X with a pre-fixed number of K clusters. In this case, the latent topics play the role of document clustering. The probability of observing the (sub)-topic or cluster c_k given the document x , $P(c_k|x)$ is interpreted as the probability that the document x belongs to the cluster c_k . In the next step, each document is assigned to a *single* cluster based on

the topic distribution given a document. In other words, a document x is assigned to a cluster c_k such that:

$$\text{cluster}(x) = \underset{c_k \in C}{\operatorname{argmax}} P(c_k|x) \quad (5.6)$$

where $P(c_k|x)$ is estimated using the PLSA or LDA model. LDA is a generalisation of PLSA, where the difference between them is that in LDA the topic distribution is assumed to have a Dirichlet prior. The PLSA model will be equivalent to the LDA model under a uniform Dirichlet prior distribution [Girolami and Kabán, 2003].

5.3.2.2 Post-Clustering Methods for Document Selection

Two typical methods can be observed for selecting documents after sub-topics estimations. One method is based on the objective function that uses a ranking criterion optimised during the sequential ranking process. At each rank position, the document that meets the ranking criterion, e.g. the highest document-cluster relevance, is selected. The approaches belonging to this method are, for example, the *interpolation approach* [Kurland and Lee, 2004], cluster language model [Carterette and Chandar, 2009], IA-select [Agrawal et al., 2009], and so forth. The other method is based on a two-stage process, which first ranks clusters and then selects documents within clusters [Deselaers et al., 2009; Halvey et al., 2009; Zhao and Glotin, 2009]. In this thesis, we select one approach from each of them to study their effectiveness for result diversification.

Interpolation approach. This approach is based on cluster-based retrieval models and thus directly inspired by the cluster hypothesis. It prescribes that the relevance estimation of a document should be interpolated with the information obtained by clusters of similar documents [Kurland and Lee, 2004]. Formally, the retrieval score of a candidate document x_i is calculated as:

$$\hat{P}(x_i, q) = \lambda P(x_i, q) + (1 - \lambda) \sum_{c_k \in C} P(c_k, q) P(x_i, c_k) \quad (5.7)$$

where c_k is a cluster of similar documents in C , i.e. the set of document clusters modelled by (sub)-topic modelling approaches; λ indicates the degree of emphasis on the probability of relevance and the probability of the document belonging to a cluster. In the context of our study, we assume that $P(a, b)$ is a similarity function between the objects¹ a and b . For example, $P(x_i, c_k)$ is simplified as the similarity between document x_i and cluster c_k , where the cluster is represented by cluster centroid vector. Note that when $\lambda = 0$, the ranking function of equation (5.7) returns documents within the cluster with highest similarity to the query, i.e. the cluster with higher $P(c_k, q)$. Furthermore, the interpolation approach employs the evidence of multiple sub-topics to rank documents. In other words, a document can belong to two or more sub-topics. The documents ranked in the early positions are those, which are most likely to be members of the clusters and those clusters are highly relevant to a query. In the empirical evaluation, we indicate this approach with **Interp(.)**.

Cluster representative approach. This approach assumes that each cluster represents a different sub-topic. Thus, to the complete array of sub-topics, cluster representatives (documents within a cluster) of every cluster have to be retrieved at early ranks. Assume that we have a reliable clustering method $Cluster(.)$ that estimates sub-topics by grouping similar documents into clusters, and a ranking method $C\text{-Ranker}(.)$ that rank clusters with respect to their relevance to a query. The approach for cluster representative selection is described as follows.

For a given set of documents X , we generate a set of K clusters using three different instances of $Cluster(.)$. We perform clustering using K-means, PLSA and LDA models as described in Section 5.3.2.1. Then, we rank clusters in decreasing order of the relevance of the clusters to a given query (e.g. $P(c_k, q)$), which results in a ranked list of clusters, i.e. $C\text{-Rank} = c_1, c_2, \dots, c_k$. For the purpose of the study, we define the $C\text{-Ranker}(.)$ method by using the average relevance of the documents contained in each cluster. Given a query q and a cluster c_k , average cluster relevance is defined as:

$$C\text{-Ranker}(.) \Rightarrow S_{avg}(c_k, q) = \frac{1}{N_k} \sum_{i=1}^{N_k} s(x_{k,i}, q) \quad (5.8)$$

¹These can be queries, documents, or clusters.

5.3 Background of Result Diversification

where N_k is the number of documents in c_k and $X = \{x_1, \dots, x_N\}$ is the initial set of relevant documents. Average cluster relevance is employed for ordering the clusters; then each cluster is selected in a round-robin fashion, following the order suggested by the average cluster relevance. That is, in each round, every cluster is ensured to be selected in the order of c_1, c_2, \dots, c_k . Within each cluster, several strategies can be employed to select an individual document as a cluster representative. This is what is different about the instantiation of different methods belonging to this approach.

For example, in [Ferecatu and Sahbi, 2008] cluster representatives are selected according to the order in which documents are added to clusters. An alternative approach is suggested by Urruty et al. [2009] where the medoid¹ is assumed to be the best cluster representative. Zhao and Glotin [2009] suggest to choose the document with the lowest rank within each cluster of the top retrieved results. Halvey et al. [2009] propose to select the document that is most similar to other members of the selected cluster. Finally, cluster representatives are selected according to the highest document relevance to a query or the original score of the retrieved document [Deselaers et al., 2009]. In our empirical study we opt to investigate the latest solution by using the *document relevance* or PRP, which we denote in the evaluation with **Repre_{PRP}**(.).

Figure 5.1(b) visualises the document selection procedure following the cluster representative approach with the round-robin algorithm. Within each cluster (e.g. grey, yellow, blue, etc.), documents (e.g. x_1, x_2, x_3 , etc.) are selected according to their highest relevance. In other words, this approach selects a document with the shortest distance to a query q . Although this approach ensures, to some extent, that documents from different clusters contain diverse contents, we argue that the strategy of document selection is based on a fixed criterion, i.e. PRP², which may lead to the redundancy of information in a ranking. That is, the selection strategy lacks the estimation of document dependencies. By following this strategy, documents at nearby positions are still adjacent to each other in a metric space (e.g. cosine similarity metric) because the selection criterion is only associated with a single point of query q (or few points of centroids μ in medoid-based selection). In particular, the documents from the same cluster (i.e. x_1 vs x_5 , x_2 vs x_6 , etc.) are chosen in close proximity. Furthermore, *not*

¹The document closest to the centroid of the cluster.

²This is not limited to other document selection strategies such as cluster centroid, the lowest probability of relevance, and so forth.

all documents in a cluster are relevant to user information needs. Therefore, if two very similar documents are retrieved and one of the document is judged non-relevant, the other document tends to be non-relevant as well. The chance of retrieving relevant documents is minimised by following this selection strategy. We suggest that the strategy of document selection can be improved for better diversification. We therefore consider the generalisation of this approach to our framework, aiming to improve its results by incorporating inter-dependent document relevance.

5.4 Diversification with Two Ranking Paradigms

In this section, we introduce our proposed framework for integrating *inter-dependent document relevance* and *sub-topic aware* paradigms. The overall goal of the integration approach is to systematically generalise/combine these paradigms into a unified framework to increase the effectiveness of diversifying documents as measured in terms of both relevance and diversity.

5.4.1 Motivation

Two ranking paradigms for result diversification have been recognised in information retrieval. First, the inter-dependent document relevance paradigm *implicitly* achieves sub-topic coverage by considering the relationship between documents using objective functions. Even though the effectiveness of retrieval systems is mainly evaluated by the number of unique relevant sub-topics, the approaches belonging to this paradigm do not have any estimation regarding the presence of sub-topics. These approaches rely on retrieving relevant documents that contain little redundant information compared to other documents, i.e. documents are different from each other and such difference may refer to previously undiscovered sub-topics. Therefore, if this paradigm is followed, documents are *not* directly diversified with respect to sub-topics and thus documents within nearby positions can be chosen from the same sub-topics (see Figure 5.1(a)). We have argued that ranking documents based on this paradigm may not be effective enough to achieve a complete sub-topic coverage at early ranks.

Secondly, inspired by cluster-based retrieval, the sub-topic aware paradigm *explicitly* estimates hypothesised sub-topics in terms of clusters from retrieved documents.

5.4 Diversification with Two Ranking Paradigms

This paradigm interleaves documents belonging to different clusters so as to cover all possible (or identified) sub-topics in a result list. A number of ranking strategies based on static criteria, e.g. the highest/lowest document relevance, medoid, etc., have been proposed for selecting documents within particular clusters. However, we have argued that none of these strategies considers the dependencies of documents. Furthermore, this ranking paradigm only depends upon the results of sub-topic estimations obtained by topic modelling methods, and if estimated sub-topics do not provide sub-topic evidences corresponding to information needs, selecting documents from different clusters will not effectively produce the results with respect to the needs.

Therefore, we aim to investigate whether a new effective ranking approach can be devised to effectively promote diversity in a ranking. We proposed the integration framework based on two ranking paradigms. The motivation behind this framework is as follows.

- By clustering documents, topically coherent groups of documents that encode possible sub-topics of a general topic are formed.
- Although hypothetical partitions of sub-topic are estimated, the diversity of a ranking relies solely on the correctness of sub-topic modelling. If estimated sub-topics do not corresponds to the users common perception of sub-topics, an inter-dependent document ranking paradigm could assist in finely tuning document rankings to cover relevant sub-topics for the users.
- With sub-topic estimates, existing document selection methods *discard* either the document relevance, document dependencies, or both, implying the causation of redundancy in document ranking. The redundancy of documents can be alleviated by including document dependencies during document selection.

5.4.2 Proposed Framework

In Figure 5.2, we illustrate the proposed diversification framework that integrates sub-topic clusters and document dependencies in a result list. Within the framework, we propose a two stage procedure developed for diversification. In the first stage, we

5.4 Diversification with Two Ranking Paradigms

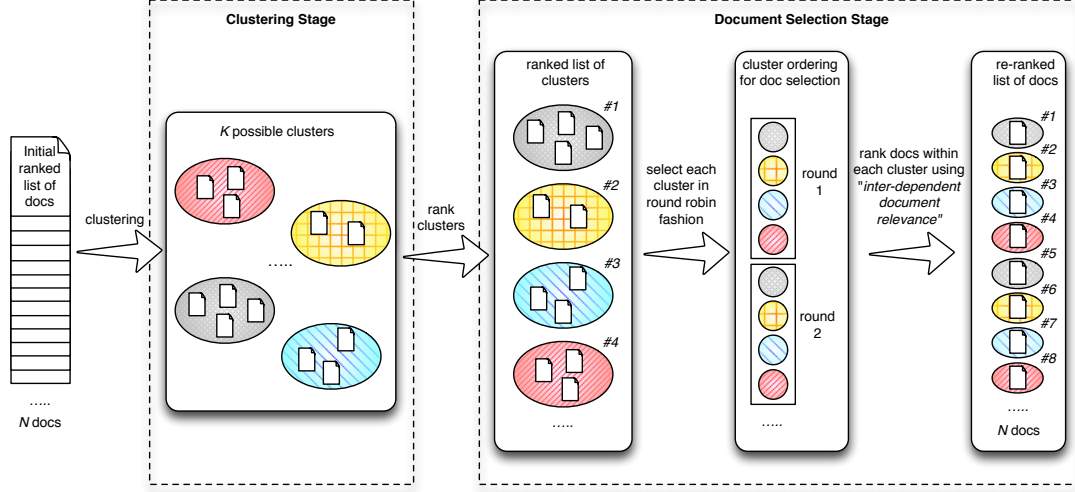


Figure 5.2: Diversification with cluster ranking. The input is a ranked list of documents and output is a diversified ranked list of documents.

assume that there is the reliable clustering method $Cluster(.)$ that estimates hypothesised sub-topics from a set of initially retrieved documents X . We employed K-means, PLSA, and LDA models as the $Cluster(.)$. In the set of documents X we partition documents into the pre-fixed number of K clusters, where each document is assigned to a single cluster (see Section 5.3.2.1). The given sub-topic estimates are then used for document selection in the next stage.

The goal of the second stage is to produce a re-ranked list of documents with the inclusion of document dependencies in selecting documents within a cluster. Assume that we have a cluster ranking method $C-Ranker(.)$ that ranks clusters in decreasing order of cluster relevance to a query, and an inter-dependent document relevance method $Inter-Dep(.)$ that diversifies a ranked list of documents by accounting for the dependency of documents. $C-Ranker(.)$ is assumed to be an average of document relevance within a cluster as computed by equation (5.8). We then obtain a ranked list of clusters $C-Rank = c_1, c_2, \dots, c_k$, where the documents contained in each cluster c_k is denoted as $x_k \in c_k$. Sub-topic clusters are subsequently selected in a round-robin fashion according to the order of $C-Rank$. That is, in each round, we apply $Inter-Dep(.)$ to select an individual document from each of the clusters, and append them to a new ranked list of documents. At each ranked position, a document that maximises the method

Inter-Dep(.), is selected into the result list. This selection procedure continues until no documents are left in any of the clusters or the maximum number of documents required for retrieval is achieved.

The most important component of our proposed framework is the function *Inter-Dep(.)*. This function ranks documents based not only on their relevance to a query, but also on the relationship amongst documents in a result list. As for *Inter-Dep(.)*, we discuss our choices for the instantiation of integration approach in the following section.

5.4.3 Integration Approach

As we discussed above, ranking documents based on their dependencies is an important issue, which has been studied in the context of sub-topic retrieval. Since our main purpose is not to develop a new ranking method based on inter-dependent document relevance, we only discuss the MMR approach as our first instantiation of *Inter-Dep(.)* for a family of integration approaches.

Formally, if the MMR-like function is used as *Inter-Dep(.)*, then the following objective function should be maximised:

$$J_j = J_{j-1} \cup \underset{x_{k,n} \in X_k \setminus J}{\operatorname{argmax}} [\lambda S(x_{k,n}, q) + (1 - \lambda) \underset{x_j \in J}{\operatorname{avg}} D(x_{k,n}, x_j)] \quad (5.9)$$

where $X_k = \{x_{k,1}, x_{k,2}, \dots, x_{k,n}\}$ is the set of retrieved documents that belong to the sub-topic cluster c_k and J is the set of documents that has already been ranked. Of course, other approaches, such as PT, qPRP, or iPRP, can be used to replace MMR. However, we restrict the scope of our investigation to MMR so as to investigate the effectiveness of our proposed framework for result diversification.

The pseudocode of our diversification with integration approach is outlined in Algorithm 2: this is the same algorithm that has been implemented to produce the results reported in our empirical investigation (Chapter 6). The algorithm applies *Inter-Dep(.)*, i.e. MMR function, to the documents within the cluster that is selected using the round robin algorithm. Following the criterion of MMR, at rank position j the document that is in the selected cluster c_k and contains the highest marginal relevance (i.e. relevant information that is not similar to that already presented) is selected. The hyper-parameter

5.4 Diversification with Two Ranking Paradigms

Algorithm 2 Inter-dependent document relevance (using MMR) on the evidence induced from sub-topic clusters.

Require: q , a user query

Require: $C\text{-Rank} = \{c_1, c_2, c_3, \dots, c_k\}$, a set of k clusters ranked according to $S_{avg}(c_k, q)$

Require: $X_k = \{x_{k,1}, x_{k,2}, x_{k,3}, \dots, x_{k,n}\}$, a set of n documents in cluster c_k

Require: $j = 0$, where j is the number of documents that has been already ranked

Require: $maxDocs$, the maximum number of required documents to retrieve

$J_0 = \{\}$

while $j \leq maxDocs$ **do**

if $j = 0$ **then**

$J_0 = \underset{x_{k,n} \in X_k \setminus J}{\operatorname{argmax}} [S(x_{k,n}, q)]$

else

$Inter\text{-}Dep(.) \Rightarrow J_j = J_{j-1} \cup \underset{x_{k,n} \in X_k \setminus J}{\operatorname{argmax}} [\lambda S(x_{k,n}, q) + (1 - \lambda) \underset{x_j \in J}{\operatorname{avg}} D(x_{k,n}, x_j)]$

end if

$j = j + 1; k = k + 1$

if $k \geq j$ **then**

$k = 0$

end if

end while

return $J_j = \{x_1, x_2, x_3, \dots, x_j\}$, a set of re-ranked documents to present to the user

λ is used to tune the function so as to optimise the retrieval effectiveness as measured by evaluation measures, e.g. sub-topic recall or α -nDCG. In the empirical evaluation, we denote the integration approach incorporating MMR for ranking documents with $\mathbf{Integr}_{MMR}(\cdot)$.

Figure 5.1(c) depicts the results of our integration approach, in which documents x_1, x_2, \dots, x_8 are selected according to both particular estimated sub-topics and document dependencies. In each iteration, documents are selected from different clusters and so presumably from different sub-topics. Within each sub-topic, we maximise the chance of retrieving relevant documents by selecting novel documents (containing low redundant information) that are different from other documents in a ranking. For example, at rank 5 our approach selects document x_5 based on not only its relevance to a query (red line), but also its dissimilarity to documents x_1, \dots, x_4 that have been ranked previously (blue line). Specifically, documents x_1 and x_5 are selected from different areas within the same cluster (grey). This example is compared with x_1 and x_5

of Figure 5.1(b) that are selected in close proximity.

5.5 Summary

In this chapter we have discussed automatic methods for sub-topic retrieval, which model dependent relevance and promote document diversity in order to find relevant documents that cover as many sub-topics as possible. We have first presented the ranking problem of the classic information retrieval methodology, the probability ranking principle (PRP). The key assumption of PRP is the independence of document relevance, in which each of the documents is assessed individually for its relevance. Although this assumption plays a central role in the development of the information retrieval field, it often does not hold. In particular, this is the case in the context of sub-topic retrieval where the problem is concerned with dependent document relevance. Therefore, the independence assumption is critical for PRP's optimality since the usefulness of a relevant document to a requester is not defined independently, but instead depends on the *number* of relevant documents sharing the same information the requester has already seen ("the more he has seen, the less useful a subsequent document may be" [Robertson, 1977]).

Then, we have reviewed two main ranking paradigms of diversifying documents for sub-topic retrieval. We have first outlined the inter-dependent document relevance paradigm, which promotes sub-topical diversity by including dissimilarity estimation in objective functions: maximal marginal relevance, and modern portfolio theory. Other approaches have been also proposed in the IR literatures. However, we have argued that these approaches do not explicitly estimate sub-topics. Thus, if this paradigm is followed, documents within nearby positions can be selected from the same sub-topic, resulting in an unsatisfactory performance in terms of sub-topic coverage. We have secondly presented the sub-topic aware paradigm, which explicitly model sub-topics by clustering similar documents: k-means clustering, latent Dirichlet allocation, and probabilistic latent semantic analysis. We have also argued that the current methods for selecting documents after clustering do not use inter-document similarity for diversification purposes to good effect. Therefore, documents selected in different rounds of cluster selection may contain too much similar information, which

minimises the probability of returning a relevant document belonging to a particular sub-topic.

In summary, we have proposed a general result diversification framework, which enables the development of a variety of algorithms for integrating a query-specific similarity structure modelled via clusters and inter-dependent document relevance. The overall goal of our integration approach is to increase, from the use of individual paradigm, the effectiveness of diversifying documents as measured in terms of both relevance and diversity. We posit that such an integration approach should lead to improved results since sub-topics are explicitly estimated and documents are dependently ranked. Although our proposal is motivated by the classic re-ranking approach to sub-topic retrieval, i.e. MMR, we suggest that the framework can be used with other recent approaches based on the inter-dependent document relevance paradigm such as MPT, qPRP, and iPRP. Moreover, we are interested in comparing the effectiveness of the state-of-the-art approaches and our proposal for result diversification. In the following we list some issues for investigation and follow-up in relation to the development of our framework.

- Which paradigm or approach delivers the best document ranking for sub-topic retrievals?
- What is the impact of diversification when using existing result diversification approaches individually or together? In other words, how much performance is gained by integrating the diversification approaches?
- How does the variation of pairs of sub-topic modelling and inter-dependent document ranking in the integration approach affect the evaluation of document rankings?
- What are the circumstances in which the integration approach gives the best overall performance? In particular, given that estimated sub-topics correspond to user information needs, how sensitive is the performance of the integration approach to the number of relevant sub-topics being retrieved?

In the next chapter, we shall further investigate the effectiveness of different ranking approaches using the data of ImageCLEF 2009, TREC ClueWeb 2009, and TREC

6,7,8 interactive collections. We conduct a number of experiments to empirically validate and contrast the state-of-the-art approaches as well as instantiations of our integration approach.

Chapter 6

Empirical Study of Ranking Diversification for Sub-topic Retrieval

6.1 Introduction

In the previous chapter, we analysed and discussed the current state-of-the-art approaches, producing a ranking list of documents that cover as many sub-topic as possible. From these approaches, two common patterns have been observed, with respect to modality, used to produce document ranking diversification. With an aim to provide a better ranking for sub-topic retrieval task, we formalised a new general ranking framework that provides an insight to the development of a variety of algorithms, regardless of the specific choice of the similarity functions, the document dependency functions, and the sub-topic modelling algorithms. The framework is on the basis of the integration of the two ranking patterns that 1) estimate possible sub-topics covered by documents, and 2) rank documents using inter-dependent document relevance. To the best of our knowledge, no empirical study has been performed comparing and integrating them in the context of sub-topic retrieval.

This chapter presents an empirical experiment, aiming to examine which paradigm, and in turn which approach, performs the best in sub-topic retrieval task. In order to improve the effectiveness of result diversification, we evaluate whether sub-topic estimates provided by sub-topic modelling techniques (e.g. clustering, LDA, PLSA) can be employed together with inter-dependent document ranking. Furthermore, we aim to study under what condition the integration of two paradigms yield improved

performance when compared with existing approaches. To this end, we generate sub-topic clusters that correspond to sub-topic relevance judgements so that we can assume the situation when sub-topic modelling techniques can relevantly estimate sub-topics from search results. We analyse and evaluate a number of ranking approaches using various diversity measures. The experimental results are reported and discussed in this chapter.

This chapter is organised as follows. In Section 6.2, we outline the experimental plan, scope of the study, and research questions to be answered in this study. Subsequently, we present the results from the studies, followed by the discussion of their analysis in Section 6.4. This chapter concludes in Section 6.5, where we summarise the obtained results and our contributions.

6.2 Experiment and Validation

In the following subsections, we describe experimental methodology based on system-oriented evaluation. We first define the research questions that our studies want to address. Then, we define assumptions that will guide and scope the experiments. Finally, we outline the experimental plan so as to ensure collected data will adequately answer the questions.

6.2.1 Research Questions

In order to understand the performance difference amongst diversification approaches, we define a number of research questions needed to be addressed. Specifically, our experiment's aim is to answer the following research questions:

- **RQ1:** Which ranking approaches, in particular of which paradigms, yield the highest retrieval performance with respect to the ranked documents in terms of diversity measures? Is the obtained performance consistent for all different test collections?
- **RQ2:** Can different approaches or paradigms be used together for a better ranking diversification? How does the outcome of the integration approach compare with that of state-of-the-art approaches?

- **RQ3:** How much performance is gained or lost by integrating two ranking paradigms? Are there similar trends in terms of performance changes when applying inter-dependent document ranking to the framework using different sub-topic modelling techniques?
- **RQ4:** How does the variation of approaches integrated in the framework affect the evaluation of document rankings?
- **RQ5:** Under what conditions does the integration approach obtain the best overall performance of diversification? In particular, given that estimated sub-topics correspond to relevance judgements or user information needs, does the performance of the integration approach increase from that of sub-topic aware paradigm?

6.2.2 Experimental Assumptions and Scope of the Study

Since our experiments involve a number of diversification approaches and our ranking framework is the first attempt to integrate two ranking paradigms, we define some assumptions that allow us to instantiate such approaches for the purpose of evaluation. Despite limiting the scope of experiments, these assumptions are in accordance with other studies used to instantiate other ranking approaches. We believe that these assumptions can be relaxed for further investigation but here we intend to discover the preliminary benefits of the framework in bringing together two retrieval patterns. In the following, we list the experimental assumptions:

- 1) In our framework, we assume the independence amongst sub-topic clusters of documents (i.e. a document can belong to only one cluster.). This is in line with the cluster representative approaches on which our framework is based [Dese-laers et al., 2009; Ferecatu and Sahbi, 2008; Zhao and Glotin, 2009].
- 2) We employ an *interpolation* approach of sub-topic aware paradigm to examine the performance of the diversification approach, which allows for overlapping between sub-topic clusters in its objective function¹ (i.e. a document can belong to one or more clusters); and

¹This is different from the assumption made for constructing sub-topic clusters.

Table 6.1: Statistics of three experimental collections.

Name	Document		
	# Docs	# Term	# Uniq. Terms
ImageCLEF09	498,036	26,851,686	261,517
ClueWeb09-B	50,220,423	43,944,388,555	95,241,866
Trec-6,7,8	210,158	84,319,767	223,238

- 3) There is supposed to be a prior knowledge regarding the number of sub-topics in each search topic. By assuming this, sub-topic modelling techniques, such as K-means clustering, can estimate sub-topic clusters, which correspond to the actual number of sub-topics. Note that in practice this information can be obtained by, for example, query log analysis, or disambiguation pages from Wikipedia¹.

6.2.3 Plan of Experiments

This section describes the experimental setup that was used to conduct our experiments. We adopt a system-oriented approach, which means that the experiments require test collections, queries, and relevance judgements made for each document whether it is relevant or non-relevant. This approach assumes that users only pose queries (without further interactions) and want returned results which are only relevant to their queries. For each query, the retrieved results are compared with the judged documents so as to pose a statement about the performance of retrieval methods/systems.

In the following, we detail the document collection, the topics and the measures used in our evaluation. Additionally, we describe the baseline with which our approach is compared and parameter settings for all diversification approaches.

6.2.3.1 Test Collections and Topics

In order to answer research questions, we evaluate the effectiveness of state-of-the-art approaches belonging to two paradigms and our integration approach. Our experiments rely on three standard test collections for sub-topic retrieval tasks: ImageCLEF

¹<http://en.wikipedia.org>

Table 6.2: No. of topics and statistics of sub-topic in three collections.

Name	# Topics	# Sub-topics		
		Min.	Max.	Avg.
ImageCLEF09	50	1	10	3.96
ClueWeb09-B	50	1	6	4.1
Trec-6,7,8	20	7	56	20

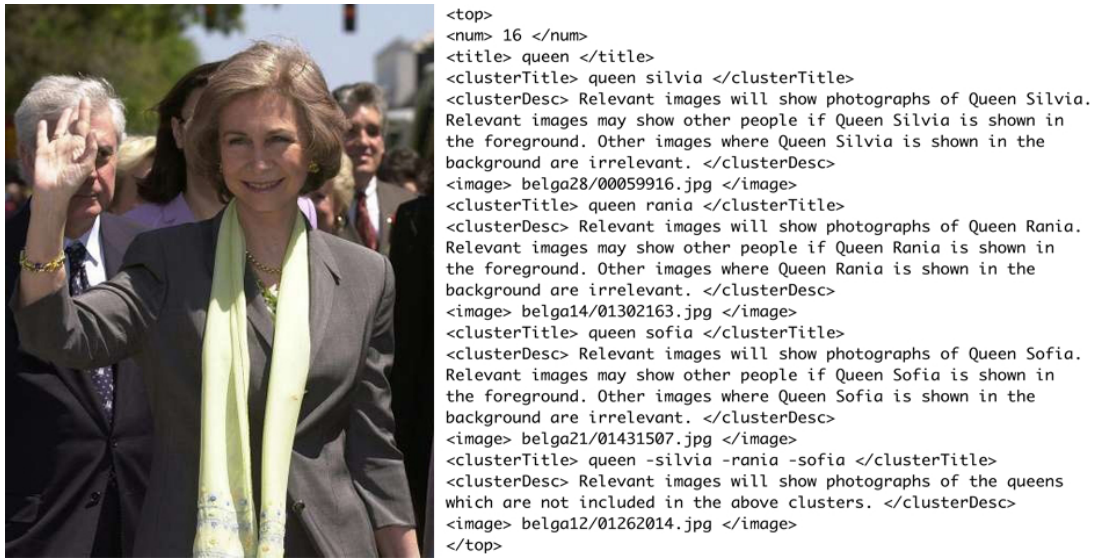


Figure 6.1: Example of ImageClef 2009 Photo Retrieval dataset entry, topic 16, with relevant image (left) as well as query topic and its sub-topics (right).

2009, TREC ClueWeb 2009, and TREC 6,7,8 interactive collections. Some statistics regarding document datasets, topics and sub-topics are shown in Table 6.1 and 6.2.

As one of our test collections, we employ ImageCLEF 2009 dataset¹, consisting of almost 500 thousand image documents from the Belga News Agency, an image search engine for news photographs. The dataset was employed as a test collection at the (diversity) photo retrieval task of ImageCLEF 2009 track [Paramita et al., 2009]. Each image is accompanied by a caption composed of English text with an average length of 53 words. Most words in the caption are keywords that indicate the content of images

¹<http://www.imageclef.org/2009/photo/>

```
<topic number="1" type="faceted">
  <query>obama family tree</query>
  <description>Find information on President Barack Obama's family
  history, including genealogy, national origins, places and dates of
  birth, etc.
</description>
  <subtopic number="1" type="nav">
    Find the TIME magazine photo essay "Barack Obama's Family Tree".
  </subtopic>
  <subtopic number="2" type="inf">
    Where did Barack Obama's parents and grandparents come from?
  </subtopic>
  <subtopic number="3" type="inf">
    Find biographical information on Barack Obama's mother.
  </subtopic>
</topic>
```

Figure 6.2: Example of TREC ClueWeb 2009 dataset, topic 1, along with its corresponding sub-topics.

well. In this experiment, we use only text caption for indexing documents and discard visual features extracted from images.

Although the collection contains only image documents, we believe that the ImageCLEF dataset is still suitable for our study. This is because our study aims to evaluate different ranking approaches, but not to evaluate data features used in retrieval systems. Furthermore, the task of ImageCLEF 2009 (i.e. topics) focuses on topical, instead of, visual diversity. As shown in Figure 6.1, an example of topic 16 is “queen”, which is topically related to sub-topics (i.e. cluster titles) such as “queen silvia”, “queen rania”, “queen sofia”, and “other queens”. Thus, there is no major concern in ruling out visual features from the empirical investigation. We use the set of 50 available topics, each of which comes with relevance judgements made at sub-topic level. Retrieval systems receive only topic titles as inputs, representing short and ambiguous queries issued by users. The cluster titles, cluster descriptions, and image examples are *not* included for querying search results.

Our analysis is also conducted using the standard experimental collection provided by the diversity task of TREC 2009 Web track [Clarke et al., 2009a]. As the underlying collection, we consider the category-B ClueWeb 2009 dataset¹, consisting of 50 million English Web documents crawled between January and February 2009. For our

¹<http://lemurproject.org/clueweb09/>

```

Number: 352i

Title: British Chunnel impacts

Description: Impacts of the Chunnel - anticipated or actual - on the British
            economy and/or the life style of the British

Instances: In the time allotted, please find as many DIFFERENT impacts of
            the sort described above as you can. Please save at least one
            document for EACH such DIFFERENT impact.
            If one document discusses several such impacts, then you need
            not save other documents that repeat those, since your goal
            is to identify as many DIFFERENT impacts of the sort described
            above as possible.

Topic
|
|      Instance#
|      | Instance gloss
|      | |
|      | |

352i    1 environmental impact
352i    2 financing of high-speed rail line
352i    3 cost of additional safety standards
... ...
352i    28 removes psychological barrier of Channel

```

Figure 6.3: Example of TREC 6, 7, 8 interactive dataset, topic 325i, along with its corresponding sub-topics, called instances.

queries, we use the titles of 50 topics from the diversity track. For each topic, there are 1 to 6 sub-topics, extracted from query logs. Relevance judgements were made with respect to each sub-topic. Figure 6.2 illustrates an example topic with different fields, including its identified sub-topics. Similar to ImageCLEF dataset, we only uses the “query” field of the topic as an initial query in our experiment. Note that at the time when conducting the experiment, only 2009’s relevance judgements¹ were publicly available for diversity evaluation on a Web setting.

Finally, we use the dataset, derived from the TREC interactive track (TREC-6, TREC-7, and TREC-8) [Hersh and Over, 2000; Over, 1998, 1999]. With the introduction of sub-topic retrieval task, Zhai et al. [2003] originally used the Financial Times of London 1991-1994 collection (part of TREC-6,7,8 ad-hoc collection) for the purpose of diversity-based experiment. By following Zhai et al.’s practice, we conduct our experiment on this test bed dataset, in which instances (i.e. sub-topics) are identified by TREC accessors from topics’ narrative section. For example, for the sample topic 325i

¹We ignored the query set from TREC 2010 since it was not available during experimentation.

they identified 28 different sub-topics, some of which are shown in Figure 6.3. This collection contains over 210 thousand documents with over 84 million terms. We used all 20 topics collected for 3 years (i.e. TREC 6, 7, 8). For each topic, the judgement for each document can be represented as a bit vector with up to 56 bits, each indicating whether the document covers the corresponding sub-topics. Similar to the other two datasets, topic titles are only used as initial queries to retrieve the initial result sets for document re-ranking.

6.2.3.2 Evaluation Measures

For evaluation, we employ three official measures used in TREC Web Diversity and ImageCLEF Photo Retrieval tracks: *sub-topic recall* (*s-recall*) [Zhai et al., 2003], *sub-topic mean reciprocal rank* (*s-mrr*) [Chen and Karger, 2006], and *novelty-biased cumulative gain* (α -*nDCG*) [Clarke et al., 2008]. We first use *s-recall*, which evaluates the diversity of relevant information in terms of sub-topic coverage in a document ranking. Specifically, *s-recall* is defined as the proportion of sub-topics covered by documents up to a given rank r with respect to the total number of sub-topics associated with a query. In addition, the diversity performance of our approach is reported in terms of *s-mrr*, defined as the inverse of the rank at which a specific percentage of sub-topic coverage is achieved (i.e. 25%, 50%) [Chen and Karger, 2006].

Besides *s-recall* and *s-mrr*, we use α -*nDCG* measure, which addresses both relevance and redundancy, and balances them through the tunable parameter α . The larger the value of α , the greater the discount applied to documents containing redundant sub-topics. In contrast, when $\alpha = 0$, only relevance is taken into consideration, and this measure is equivalent to the traditional *nDCG* [Järvelin and Kekäläinen, 2002].

Note that all three measures are thoroughly investigated in Part IV, where we propose an approach for setting the value of parameter α for α -*nDCG*. In this chapter we however follow the standard evaluation practice in TREC Web Diversity track [Clarke et al., 2009a], i.e. we compute α -*nDCG* with $\alpha = 0.5$, which give equal weights to both relevance and redundancy. The results will be re-analysed with our newly proposed setting, and presented in Chapter 9.

6.2.3.3 Experimental Systems and their Settings

– *Settings for Retrieval.*

To empirically contrast result diversification approaches, we use only textual information in our experiments. Three document collections are indexed using Lemur¹, which also serves as platform for developing the ranking approaches using C++. We remove standard stop-words [van Rijsbergen, 1979] and apply Porter stemming to both documents and queries. For each retrieval topic, a query is extracted from the title of the TREC and CLEF topics.

We use the Okapi BM25 to estimate document relevance with respect to a query, where its parameters are set according to standard values [Robertson et al., 1996]. Once estimates of document relevance are obtained using Okapi BM25, we produce an initial ranking according to the Probability Ranking Principle, i.e. we order documents with respect to decreasing probability of relevance. We denote this run with **PRP**; this represents the naive baseline in our experiments, i.e. a method without diversification. Furthermore, the BM25 weighting scheme is also used to produce document term vectors, that are used by diversification approaches for computing similarity (e.g. in MMR), correlation (e.g. in MPT), or cluster (e.g. in K-means). By doing this, our experiment is consistent with previous works [Wang and Zhu, 2009].

For each query in our test collections, we experiment with several ranking lengths, i.e. 100, 200, 500, and 1000, meaning that all the documents retrieved at ranks lower than these thresholds are discarded. In order to promote diversity, the initial rankings of documents obtained by PRP is then used as input of diversification approaches for re-ranking documents. However, in this thesis we report results for ranking up to 100 documents long since other results obtained with different ranking depths present similar results and trends.

– *Settings for Diversification.*

To obtain the similarity between *document* and *query* that is involved in diversification algorithms (e.g. MMR, MPT), we employ BM25 score, normalised into the range [0,1]. Since the BM25 retrieval score is in the log domain, we transform it back

¹<http://www.lemurproject.org/>

into the original domain. For each document x in the initial result set I , we normalise the document BM25 score using:

$$\text{norm}(S(x, q)) = \frac{S(x, q)}{\sum_{x_i \in I} S(x_i, q)}$$

We can then define the parameters of diversification approaches as follows:

Maximal Marginal Relevance (MMR). We instantiate the MMR approach as discussed in Section 5.3.1.1, where we employ the normalised BM25 score as similarity function between document and query, and the opposite of the cosine similarity between documents as a measure of dissimilarity. Furthermore we vary the value of λ in the range $[0,1]$ with steps of 0.1.

Modern Portfolio Theory (MPT). When testing MPT, we explore values of b , the risk propensity, in the range¹ $[-9, 9]$; we treat the variance of a document as a parameter that is constant with respect to all the documents, similarly to [Wang and Zhu, 2009]. We experiment with variance values δ^2 ranging from 10^{-9} to 10^{-1} , and selected the ones that achieve the best performances in combination with the values of b through a grid search of the parameter space. Correlation between documents is computed by the Pearson’s correlation between the term vectors representing documents.

Sub-topic Aware Paradigm. Regarding the runs based on the sub-topic aware paradigm, we adopt three techniques to model sub-topics of documents: K-means clustering, PLSA and LDA, although alternative strategies may be suitable. The three sub-topic modelling methods are performed as described in Section 5.3.2.1. For each query, the number of sub-topic clusters required by the techniques has been set according to the sub-topic relevance judgements provided by test sets for that query. When techniques like LDA and PLSA are used, we obtain an indication of the probability that a sub-topic is covered by a document. Because in our study we do not consider overlapping clusters of sub-topics, we only assign one sub-topic to each document: i.e. the sub-topic that has been estimated as the most likely for that document. After the sub-topic

¹Note that when $b = 0$ the ranking of MPT is equivalent to the one of PRP.

clusters are formed, documents are ranked according to two state-of-the-art approaches we illustrated in Section 5.3.2.2, specifically:

- **Interp(.)**: selects documents that maximise the interpolation algorithm for cluster-based retrieval;
- **Repre_{PRP}(.)** : selects documents with the highest probability of relevance in the given sub-topics;

Interp(.) requires to build a vector which represents the cluster in order to compute $\text{sim}(c, q)$, $\text{sim}(c, d)$, and the distance to the centre of the cluster. To this aim we create cluster's centroid vector: for a cluster c_k the cluster representative vector is expressed by $\vec{c}_k = (\bar{w}_{1,k}, \bar{w}_{2,k}, \dots, \bar{w}_{t,k})$, where $\bar{w}_{t,k}$ is the average of the term weights of all the documents within cluster c_k . Cosine similarity is used to evaluate the similarity of clusters against query and document.

Repre_{PRP}(.) does not require parameter tuning. On the contrary, when instantiating Interp(.), we vary its hyper-parameter in the range [0,1] to perform optimisation with respect to diversification with entire ranked lists. The combinations of the sub-topic estimation algorithms and the document selection criteria form six experimental instantiations in total that we tested in our empirical study, i.e. **Interp(K-Mean)**, **Repre_{PRP}(K-Mean)**, **Interp(PLSA)**, **Repre_{PRP}(PLSA)**, **Interp(LDA)**, and **Repre_{PRP}(LDA)**.

Integration Approach. Following the sub-topic aware paradigm, estimated sub-topics are discovered by K-means clustering, PLSA and LDA. We then apply MMR ranking function as the first instantiation of the integration approach (see the Section 5.4). Similar to MMR's parameter setting, we vary the value of λ , ranging from 0 to 1 with steps of 0.1 and select the value that obtained the best performances in terms of diversity measure. In our experiment, the integration approach is indicated with:

- **Integr_{MMR}(.)**: selects documents according to MMR, as an example of strategy based on the inter-dependent document relevance paradigm.

With the combinations of sub-topic modelling techniques, our integration approach forms three experimental instantiations in total, i.e. **Integr_{MMR}(K-Mean)**, **Integr_{MMR}(PLSA)**, and **Integr_{MMR}(LDA)**.

Generate the Ideal Sub-topics. In addition to the use of sub-topic estimation techniques, we investigate the situation where sub-topic coverage evidence is drawn from the relevance judgements. We maintain the same assumption that a document can cover only one sub-topic: although this assumption is restrictive (and not true), it is consistent to sub-topic aware paradigm and adequate in the context of our study¹. In relevance judgements, documents that have been judged as belonging to *only* one sub-topic are assigned to a specific cluster that represents the sub-topic. These documents are then used to construct clusters' centroid vectors by averaging their term weight vectors. Afterwards, we use the Euclidean distance metric to assign those documents that cover two or more sub-topics to the closest cluster. The clusters' centroid vector are then updated with newly added documents.

The documents, which have been judged as non-relevant or do not exist in relevance judgements (also assumed as non-relevant documents), are assigned to clusters using the same procedure but are compared to all possible clusters of the topic. By doing so, we forms the sub-topic clusters containing both relevant and non-relevant documents. Therefore, we can examine the benefits of integration approach in finding at least one relevant document in sub-topic clusters. The instantiations of the approaches based on this sub-topic evidence (denoted by “**Ideal Sub-topics**”) are an indication of the upper bound performances each approach can achieve.

6.3 Results and Analysis

In this section we present the results obtained by the instantiations of the ranking strategies considered in our empirical investigation. The results are reported in Tables 6.3, 6.4, 6.5 for ImageCLEF 2009, TREC ClueWeb 2009, and TREC 6,7,8 interactive collections respectively. Results are evaluated using α -nDCG, s-recall (s-r), and s-mrr. Regarding the parametrisation of some approaches, we report here only the *best* results of each ranking approach, optimised with respect to the *average* of α -nDCG@10 when $\alpha = 0.5$ over the complete set of topics in each dataset². Parameter values are shown underneath the methods. Note that the results report here are not considered as

¹Further work will be directed towards a methodology for generating sub-topic clusters where this assumption is relaxed.

²The results produce the highest average score of all topics.

the upper bound results of *oracle* runs, where their parameters are tuned on query by query basis. However, by reporting the results of tunable runs optimised by averaging over the set of topics, we consider this to be a fair comparison against the runs without parameter tuning.

The runs of the integration approach upon different combinations of their base methods are underlined in the tables. In addition, the results that are highlighted in bold show the best performance of the runs regarding the given measures. On top of that, we report the results based on *Ideal Sub-topics*, representing the upper bound performance each technique can achieve. When statistical significant differences (according to a two-tailed t-test, with $p < 0.05$) against MMR and MPT are individuated, we report them with ^{*} and [†] respectively. We compute statistical significance against MMR and MPT because they are considered as state-of-the-art approaches in the context of sub-topic retrieval. Note that for the results of TREC 6,7,8 interactive dataset shown in Table 6.5, the statistical significance analysis is not reported, as the number of topics is very limited (just 20 topics); thus calculating statistical significance does not convey meaningful information [Bartlett et al., 2001; Voorhees and Harman, 2005].

6.3.1 Results in ImageCLEF 2009

The results we obtain on the ImageCLEF 2009 test collection suggest that instantiations of our integration approach, $\text{Integr}_{MMR}(\cdot)$, outperform those of the inter-dependent document relevance paradigm (i.e. MMR and MPT), with respect to α -nDCG@10 and when sub-topics are estimated using LDA. Other sub-topic estimation techniques (i.e. PLSA and clustering) obtain comparable results. In particular, the best results overall (at least when considering¹ α -nDCG@10) are obtained by our integration approach using LDA for estimating sub-topics, $\text{Integr}_{MMR}(\text{LDA})$. Although the performance difference of $\text{Integr}_{MMR}(\text{LDA})$ is only statistically significant against MPT, it always shows better performance over MMR. Thus integrating the two retrieval paradigms improves performances in the case of ImageCLEF 2009.

In comparison of two ranking paradigms, we see that the inter-dependent document relevance paradigm, in particular MMR, outperforms the sub-topic aware paradigm, i.e. the interpolation approach, $\text{Interp}(\cdot)$ and the cluster representative approach,

¹Note that parameters have been tuned according to this measure.

6.3 Results and Analysis

Table 6.3: Retrieval performances on the *ImageCLEF 2009 (Photo Retrieval)* collection with % of improvement over PRP. Parametric runs are tuned w.r.t. α -nDCG@10 ($\alpha = 0.5$). Statistical significances at 0.05 level against MMR, and MPT are indicated by * and † respectively.

		Models	α -nDCG@10	s-r@10	s-r@20	s-mrr 25%	s-mrr 50%
		PRP	0.4550	0.5330	0.6235	0.7589	0.5221
		MMR ($\lambda = 0.7$)	0.4830 (+6.15%)	0.6651 (+24.80%)	0.7315 (+17.33%)	0.7297 (-3.85%)	0.5041 (-3.44%)
		MPT ($b = 4, \delta^2 = 10^{-1}$)	0.4450* (-2.20%)	0.5648* (+5.97%)	0.6636* (+6.44%)	0.7307 (-3.72%)	0.4916 (-5.84%)
Sub-topic Estimation	K-means	Interp ($\lambda = 1.0$)	0.4550 (0.00%)	0.5330* (0.00%)	0.6235* (0.00%)	0.7589 (0.00%)	0.5221 (0.00%)
		Repre _{PRP}	0.4660 (+2.42%)	0.5701* (+6.97%)	0.6573* (+5.43%)	0.7503 (-1.13%)	0.5173 (-0.92%)
		Integr _{MMR} ($\lambda = 0.9$)	0.4860 [†] (+6.81%)	0.6256 [†] (+17.39%)	0.6910* (+10.83%)	0.7588 (-0.01%)	0.4985 (-4.53%)
	PLSA	Interp ($\lambda = 1.0$)	0.4550 (0.00%)	0.5330* (0.00%)	0.6235* (0.00%)	0.7589 (0.00%)	0.5221 (0.00%)
		Repre _{PRP}	0.4730 (+3.96%)	0.5766* (+8.19%)	0.6805* (+9.15%)	0.7608 (+0.25%)	0.5361 (+2.69%)
		Integr _{MMR} ($\lambda = 0.9$)	0.4950 [†] (+8.79%)	0.6520 [†] (+22.33%)	0.7179 (+15.14%)	0.7743 (+2.03%)	0.4865 (-6.81%)
	LDA	Interp ($\lambda = 1.0$)	0.4550 (0.00%)	0.5330* (0.00%)	0.6235* (0.00%)	0.7589 (0.00%)	0.5221 (0.00%)
		Repre _{PRP}	0.4740 (+4.18%)	0.5683* (+6.62%)	0.6637* (+6.45%)	0.8104* [†] (+6.79%)	0.5406 (+3.55%)
		Integr _{MMR} ($\lambda = 0.9$)	0.5020 [†] (+10.33%)	0.6236* [†] (+17.01%)	0.6842* (+9.74%)	0.7973 (+5.06%)	0.5223 (+0.04%)
Ideal Sub-topics	Interp ($\lambda = 1.0$)	0.4550 (0.00%)	0.5330* (0.00%)	0.6235* (0.00%)	0.7589 (0.00%)	0.5221 (0.00%)	
	Repre _{PRP}	0.5700 [†] (+25.27%)	0.7901* [†] (+48.24%)	0.8066* [†] (+29.37%)	0.7440 (-1.97%)	0.5544 (+6.18%)	
	Integr _{MMR} ($\lambda = 0.9$)	0.6080* [†] (+33.63%)	0.8066* [†] (+51.33%)	0.8066* [†] (+29.37%)	0.8183* [†] (+7.83%)	0.6241* [†] (+19.54%)	

Repre_{PRP}(.). However, MPT fails to improve diversification performance against Interp(.), Repre_{PRP}(.), and even the PRP baseline. Therefore, by using only ImageCLEF dataset we cannot conclude which paradigm performs the best for document diversification.

When we investigate diversification results with respect to α -nDCG and s-recall, the performance of runs based on the integration approach, Integr_{MMR}(.), consistently improve over the cluster representative approach, Repre_{PRP}(.). Note that the integration approach is inherit from Repre_{PRP}(.), where Integr_{MMR}(.) employs a round-robin algorithm to select sub-topic clusters ranked according to cluster relevance, and the

difference between them is that $\text{Integr}_{MMR}(\cdot)$ applies MMR to select documents within clusters. As the difference is the method of document selection, the evaluation results indicate that the integration with MMR increases diversification performance from the use of PRP. Although the integration approach does not perform best in terms of s-recall and s-mmrr (highlighted in bold), this may be because of two reasons. Firstly, the results reported here are not optimised with respect to s-recall and s-mmrr. Secondly, there are conflicts between α -nDCG and s-recall, and between α -nDCG and s-mmrr. We will discuss and remark the second issue in the Part IV of this thesis.

The results obtained employing evidence derived from the *ideal sub-topics* configuration indicate how much each sub-topic aware paradigm would perform if sub-topics were correctly identified. We see that in this case, the integration approach performs the best for all evaluation measures.

6.3.2 Results in ClueWeb 2009

Now let us look at the performance difference of diversification in the TREC ClueWeb 2009 dataset. Table 6.4 compares diversification with inter-dependent document relevance paradigm, sub-topic aware paradigm, and integration approach. As reported in the table, approaches based on the sub-topic aware paradigm only slightly outperform (with respect to α -nDCG@10) approaches based on the inter-dependent document relevance. In particular, this is evident when the runs obtained by MPT are compared against the runs obtained by $\text{Interp}(\cdot)$ and when the MMR runs are compared against the $\text{Repre}_{PRP}(\cdot)$ runs. However, we can notice that the performances of the sub-topic aware approaches do not much vary when considering different sub-topic estimation techniques. If the ideal sub-topic estimation is considered, then the $\text{Repre}_{PRP}(\cdot)$ approach is shown to outperform the instantiations of the other state-of-the-art approaches. Nevertheless, in this scenario our integration approach performs better than any other method, and yields up to 16.5% improvement over the $\text{Repre}_{PRP}(\cdot)$. The performance difference between the approaches that use the estimated sub-topic evidence and the ones that employ the ideal sub-topic evidence suggests that sub-topic estimation techniques fail to capture sub-topics. This might be because of the noisier nature of the ClueWeb collection relative to the ImageCLEF collection.

6.3 Results and Analysis

Table 6.4: Retrieval performances on the *TREC ClueWeb 2009* collection with % of improvement over PRP. Parametric runs are tuned w.r.t. α -nDCG@10 ($\alpha = 0.5$). Statistical significances at 0.05 level against MMR, and MPT are indicated by * and † respectively.

		Models	α -nDCG@10	s-r@10	s-r@20	s-mrr 25%	s-mrr 50%
		PRP	0.0680	0.1606	0.2719	0.1787	0.0953
		MMR ($\lambda = 0.7$)	0.1050 (+54.41%)	0.1664 (+3.65%)	0.2451 (-9.86%)	0.1741 (-2.58%)	0.0786 (-17.53%)
		MPT ($b = -5, \delta^2 = 10^{-4}$)	0.1510 (+122.06%)	0.2676* (+66.64%)	0.3486* (+28.20%)	0.2179 (+21.90%)	0.1264 (+32.69%)
Sub-topic Estimation	K-means	Interp ($\lambda = 0.2$)	0.1670* (+145.59%)	0.1682 [†] (+4.77%)	0.2331 [†] (-14.27%)	0.3411* (+90.84%)	0.1367 (+43.44%)
		Repre _{PRP}	0.1030 [†] (+51.47%)	0.1819 [†] (+13.29%)	0.2466 [†] (-9.32%)	0.2077 (+16.21%)	0.1145 (+20.21%)
		Integr _{MMR} ($\lambda = 1.0$)	0.1270 (+86.76%)	0.2019 (+25.74%)	0.2642 [†] (-2.82%)	0.2913 (+62.96%)	0.1365 (+43.31%)
	PLSA	Interp ($\lambda = 0.3$)	0.1670* (+145.59%)	0.1682 [†] (+4.77%)	0.2331 [†] (-14.27%)	0.3411* (+90.84%)	0.1367 (+43.44%)
		Repre _{PRP}	0.1160 (+70.59%)	0.1876 (+16.81%)	0.2858 (+5.10%)	0.2265 (+26.73%)	0.1120 (+17.55%)
		Integr _{MMR} ($\lambda = 1.0$)	0.1440* (+111.76%)	0.2099 (+30.72%)	0.2926 (+7.62%)	0.3140* (+75.69%)	0.1490* (+56.41%)
	LDA	Interp ($\lambda = 0.2$)	0.1670* (+145.59%)	0.1682 [†] (+4.77%)	0.2331 [†] (-14.27%)	0.3411* (+90.84%)	0.1367 (+43.44%)
		Repre _{PRP}	0.1130 (+66.18%)	0.2047 (+27.46%)	0.2902 (+6.74%)	0.2134 (+19.40%)	0.0990 (+3.93%)
		Integr _{MMR} ($\lambda = 1.0$)	0.1260 (+85.29%)	0.2149 (+33.84%)	0.2741 (+0.81%)	0.2333 (+30.51%)	0.1211 (+27.15%)
	Ideal Sub-topics	Interp ($\lambda = 0.1$)	0.1670* (+145.59%)	0.1682 [†] (+4.77%)	0.2331 [†] (-14.27%)	0.3411* (+90.84%)	0.1367 (+43.44%)
		Repre _{PRP}	0.2000* (+194.12%)	0.3332* (+107.53%)	0.3872* (+42.42%)	0.2868* (+60.48%)	0.1780* (+86.85%)
		Integr _{MMR} ($\lambda = 0.1$)	0.2330* (+242.65%)	0.3376* (+110.23%)	0.3774* (+38.81%)	0.4041* [†] (+126.09%)	0.1891* (+98.46%)

When considering all the runs of the interpolation approach, Interp(.), no matter what method is used for sub-topic estimation (i.e. K-means clustering, LDA, PLSA, and ideal sub-topic), we found that the performance of the Interp(.) runs can only reach 0.1670 as the highest score with respect to α -nDCG@10. This result suggests that although the interpolation approach yields the best diversity performance in all experimented approaches, it has the limitation that cannot exceed the maximum level. On the other hand, the integration approach, Integr_{MMR}(.), has the potential to improve the diversity effectiveness if sub-topic evidence is successfully estimated, e.g. in the case of ideal sub-topic.

Table 6.5: Retrieval performances on the *TREC 6,7,8 interactive* collection with % of improvement over PRP. Parametric runs are tuned w.r.t. α -nDCG@10 ($\alpha = 0.5$). No statistical significance is computed due to the limited number of topics.

		Models	α -nDCG@10	s-r@10	s-r@20	s-mrr 25%	s-mrr 50%
		PRP	0.4260	0.3868	0.5319	0.2877	0.1618
		MMR ($\lambda = 1.0$)	0.4260 (0.00%)	0.3868 (0.00%)	0.5319 (0.00%)	0.2877 (0.00%)	0.1618 (0.00%)
		MPT ($b = -1, \delta^2 = 10^{-1}$)	0.4330 (+1.64%)	0.3735 (-3.44%)	0.4972 (-6.52%)	0.3028 (+5.26%)	0.1643 (+1.58%)
Sub-topic Estimation	K-means	Interp ($\lambda = 1.0$)	0.4260 (0.00%)	0.3868 (0.00%)	0.5319 (0.00%)	0.2877 (0.00%)	0.1618 (0.00%)
		Repre _{PRP}	0.2380 (-44.13%)	0.2517 (-34.94%)	0.3483 (-34.52%)	0.1340 (-53.43%)	0.0692 (-57.24%)
		Integr _{MMR} ($\lambda = 1.0$)	0.2380 (-44.13%)	0.2517 (-34.94%)	0.3483 (-34.52%)	0.1340 (-53.43%)	0.0692 (-57.24%)
	PLSA	Interp ($\lambda = 1.0$)	0.4260 (0.00%)	0.3868 (0.00%)	0.5319 (0.00%)	0.2877 (0.00%)	0.1618 (0.00%)
		Repre _{PRP}	0.2580 (-39.44%)	0.3132 (-19.03%)	0.4090 (-23.11%)	0.1788 (-37.84%)	0.0688 (-57.47%)
		Integr _{MMR} ($\lambda = 0.6$)	0.2630 (-38.26%)	0.3178 (-17.84%)	0.3953 (-25.68%)	0.1797 (-37.54%)	0.0657 (-59.40%)
	LDA	Interp ($\lambda = 1.0$)	0.4260 (0.00%)	0.3868 (0.00%)	0.5319 (0.00%)	0.2877 (0.00%)	0.1618 (0.00%)
		Repre _{PRP}	0.2720 (-36.15%)	0.3078 (-20.44%)	0.4049 (-23.87%)	0.2043 (-28.99%)	0.1024 (-36.69%)
		Integr _{MMR} ($\lambda = 0.4$)	0.2820 (-33.80%)	0.3111 (-19.57%)	0.3902 (-26.64%)	0.2163 (-24.82%)	0.0989 (-38.88%)
Ideal Sub-topics	Interp ($\lambda = 1.0$)	0.4260 (0.00%)	0.3868 (0.00%)	0.5319 (0.00%)	0.2877 (0.00%)	0.1618 (0.00%)	
	Repre _{PRP}	0.5060 (+18.78%)	0.5664 (+46.41%)	0.6761 (+27.12%)	0.2898 (+0.74%)	0.1575 (-2.67%)	
	Integr _{MMR} ($\lambda = 0.9$)	0.5080 (+19.25%)	0.5692 (+47.15%)	0.6793 (+27.72%)	0.2971 (+3.28%)	0.1565 (-3.28%)	

Similar to the results obtained by the ImageCLEF collection, we observe that the $\text{Integr}_{MMR}(\cdot)$ runs always outperform the $\text{Repre}_{PRP}(\cdot)$ runs with respect to all evaluation measures. This phenomenon again suggests that with sub-topic evidences MMR can enhance the performance of diversification over PRP by including document dependencies for document selection.

6.3.3 Results in TREC Interactive 6,7,8

A similar consideration can be evidenced by the results obtained on the TREC 6,7,8 interactive collection, and reported in Table 6.5. Techniques for sub-topic estimation seem to provide unsupported evidence to the approaches of sub-topic aware paradigm,

and thus these approaches perform as well as, or worse than the PRP baseline or the inter-dependent document relevance approaches, i.e. MMR and MPT. In particular note that the results of MMR and Interp(.) are obtained when their hyper-parameter λ is set to 1, that is, when their ranking formula is equivalent to the one of the PRP baseline. Thus, the diversification components in their functions do not provide any useful evidence for promoting diversity. However, when sub-topics are estimated from the relevance judgements, as in the case of the ideal sub-topics technique, the $\text{Repre}_{PRP}(\cdot)$ and $\text{Integr}_{MMR}(\cdot)$ instantiations outperform any other approach.

For $\text{Repre}_{PRP}(\cdot)$ and $\text{Integr}_{MMR}(\cdot)$, we see that, in almost all cases, using MMR instead of PRP for document selection improves the effectiveness of document ranking diversification. Although the improvement is small, i.e. 1.87% on average for all three sub-topic estimations, one reason can be observed when considering the result of the pure MMR run, the performance of which does not improve over the PRP baseline either. This observation suggests that the MMR function may not be effective for promoting diversity in the case of TREC 6,7,8 interactive.

6.4 Findings and Discussion

This section discusses the results of our experiments that aim to answer the research questions defined in Section 6.2.1. We evaluated the state-of-the-art approaches and our proposed framework for document diversification using three test collections. We analysed results obtained by diversification approaches in order to identify which ranking paradigms provide the effective retrieval performance in terms of diversity measures.

As answers to research questions **RQ1-5**, we derived the following findings from our empirical studies:

- 1) Without considering runs of ideal sub-topic, it is likely that the inter-dependent document relevance paradigm provides better performance than the sub-topic aware paradigm. We notice that in most cases, except for the ImageCLEF dataset, the performance of diversification with MMR and MPT is higher than the Interp(.) and $\text{Repre}_{PRP}(\cdot)$. Furthermore, LDA has been shown to provide the best evidences to support diversification in sub-topic modelling techniques as runs derived from it outperforms the others, i.e. PLSA and K-means clustering.

- 2) For the integration of two ranking paradigms, we see that the performance is enhanced by applying document dependencies to sub-topic evidences for document selection. Nevertheless, the diversification effectiveness of sub-topic aware paradigm and our integration approach is restricted due to poor sub-topic evidences provided by sub-topic modelling techniques, particularly in the case of TREC ClueWeb 2009 and TREC 6,7,8 interactive. We suggest that supervised learning methods can be used to alleviate this problem by identifying sub-topics, which are more relevant to sub-topic judgements representing multi-intent information needs.
- 3) A clear pattern has been observed when considering the integration approach against the original approaches of sub-topic aware paradigm. From the results, we found that the performance of the integration approach increases or at least remains the same when compared with that of sub-topic aware paradigm. As we can see the maximum gains that our framework can potentially achieve are 33.63%, 242.65%, and 19.25% with respect to the PRP baseline for α -nDCG@10 in ImageCLEF 2009, TREC ClueWeb 2009 and TREC 6,7,8 interactive datasets respectively.
- 4) There is not much difference between results of LDA and PLSA, which outperform the results of K-means clustering. Applying MMR for document selection in our diversification framework improve an average¹ of 5.54% , 19.36%, and 1.87% over the cluster representative approach of sub-topic aware paradigm, i.e. $\text{Repre}_{PRP}(\cdot)$, in three test collections, i.e. ImageCLEF 2009, TREC ClueWeb 2009 and TREC 6,7,8 interactive, respectively.
- 5) The results of our investigation suggest that the sub-topic aware paradigm relies on the performance of sub-topic estimation techniques. This is especially evident in two TREC collections where sub-topic modelling techniques do not effectively perform in modelling sub-topics. For runs obtained by using ideal sub-topics, the integration approach increases the performance of sub-topic aware paradigm. From these findings, it is suggested that the integration approach has the potential to improve document ranking diversification.

¹It is an average over three sub-topic modelling techniques, K-means clustering, PLSA, and LDA.

6.5 Summary

In this part, we introduced a diversification framework that enables the development of various algorithms for incorporating two ranking paradigms in sub-topic retrieval. We show and highlight that our diversification framework is more flexible than other diversification approach when it comes to integration in more complex result presentation strategies. We reviewed and discussed the state-of-the-art approaches for promoting diversity in document ranking. We illustrated two different viewpoints in developing each approach, which actually aims to achieve the same goal of completing sub-topic coverage and avoiding excessive redundancy. While one is developed through dependent relevance models, the other is devised to predict sub-topics from the relationship between documents. We proposed that they can be used together so as to improve ranking diversification.

To assess the effectiveness of our framework in comparison with the state-of-the-art approaches, we conduct a thorough empirical experiment using the ImageCLEF 2009, TREC ClueWeb 2009, and TREC 6, 7, 8 interactive collections. The results of our empirical investigation suggest that overall approaches derived from the sub-topic aware paradigm perform better (and in many cases significantly better) than approaches based on the inter-dependent document relevance paradigm. Amongst the techniques for estimating sub-topics, LDA and PLSA have been shown to provide better evidences than K-means clustering. However, all the techniques for estimating sub-topics fail to some extent to provide high quality evidence in the case of the TREC ClueWeb 2009 and the TREC 6,7,8 interactive collections. This might be due to the noisy nature of the documents contained in the collections (web pages and newswire articles). The integration approach, which combines implicit and explicit approaches for ranking diversification, has been shown to outperform state-of-the-art approaches, in particular when sub-topics are directly derived from the relevance judgements. Thus, the integration approach has the capability to improve sub-topic retrieval performances when effective topic estimation is deployed.

Part IV

Evaluation Measures in Sub-topic Retrieval

Chapter 7

Diversity and Redundancy-Based Measures

7.1 Introduction

In Part III, we discussed information retrieval approaches to retrieve documents with respect to multiple sub-topics (also called query-intents¹) relevant to the user's information need. Search result diversification has gained attention as an effective means to deal with the ambiguity and uncertainty of a user's query, i.e. returning relevant documents that address all possible different interpretations of such a query. Accordingly, there has been great interest in devising effectiveness measures to compare diversification methods in the context of sub-topic retrieval.

An essential element of research in information retrieval is the design of evaluation methods that allow systematic and objective comparison between different retrieval systems. Typically, when new retrieval approaches (not only for result diversification) have been proposed, we must ask:

- How well does a retrieval approach serve a user?
- How do we know which of the retrieval strategies are effective and in which applications or contexts?
- Should search scientists deem a retrieval approach as superior due to whether it *sounds* or *feels* better than another approach?

¹Here we use the terms intent and sub-topic interchangeably.

Often ideas that are intuitively thought to be able to improve search quality have actually little or no impact when empirically evaluated using quantitative experiments. A robust evaluation can provide valuable insights into whether retrieval approaches have intuitively and effectively achieved the desired goals or not. The performance of a system can hence guide the development of new approaches for better retrieval performance.

In the context of sub-topic retrieval, the goals as defined by the TREC 2009–11 Web Diversity track guideline¹ are:

- 1) to return a diversified list of relevant documents that provide complete coverage of sub-topics given a query; and
- 2) to minimise the amount of redundancy with respect to such sub-topics.

It is of necessary importance to understand whether the measures satisfy and meet the goals of the evaluation context.

In this chapter, we first discuss the main objectives of retrieval systems developed for result diversification. Then, we outline three major evaluation measures, i.e. *sub-topic recall* (*s-recall*), *sub-topic mean reciprocal rank* (*s-mrr*), and *novelty-biased discounted cumulative gain* (α -*nDCG*). These all assess search results in terms of *diversity* and *redundancy*. We focus on these measures because they are employed in our evaluation framework and are widely used in recent research and development. Note that a wider overview of measures used in this evaluation context is given in Section 2.4.2.2. Following this, we make a clear distinction between evaluation measures by classifying them with respect to diversity and redundancy, as they address the effectiveness of retrieval systems differently. While α -*nDCG* has become more prevalent than other measures in the evaluation of sub-topic retrieval, we identify that, in particular circumstances, α -*nDCG* does *not* measure systems as specified by TREC guidelines. In fact, we observe that α -*nDCG* penalises systems that cover many sub-topics while instead it rewards those that redundantly cover only few sub-topics. We propose a formal approach to allowing α -*nDCG* to achieve the evaluation expectations. We suggest that by means of our approach α -*nDCG* can turn to be more intuitive with respect to the objectives of the diversity task. This approach is based specifically upon setting α -*nDCG* on a query by query basis [Leelanupab et al., 2011].

¹See <http://plg.uwaterloo.ca/~trecweb/2010.html> guidelines or Appendix D.

7.2 Evaluation Methodology in Diversity Task

TREC 2009–11 Web Diversity tracks have been created with the aims of investigating and evaluating the performance of systems that can retrieve relevant documents while providing *diversity* of sub-topics within the search results. Retrieval systems designed for this task should retrieve documents covering a complete array of potentially relevant sub-topics (i.e. “provide complete coverage for a query”, while “avoiding excessive redundancy”¹). By adhering to this policy, the likelihood of retrieving information relevant to each sub-topic is maximised. Each sub-topic addresses a different aspect of an information need and so they all should be retrieved.

Within this context, the set of assessments that determine the relevance of a document differs from that of ad-hoc retrieval tasks. Each topic is structured as a representative set of sub-topics, each related to a different user need or query-intent. Documents are judged with respect to sub-topics of a general topic. A set of information needs is expressed by an ambiguous or faceted query (in line with TREC assumption²). Ambiguous queries are those that have many interpretations, but users who issue such queries are assumed to be interested in only one of these interpretations. On the contrary, faceted queries are those that are *underspecified* to a particular aspect of interest. It is assumed that users issuing these queries would be interested in one aspect, but may still be interested in others as well.

Once a test collection that consists of documents, queries, and sub-topic relevance judgements is available, system-centred evaluation (often referred to as *Cranfield Paradigm*) can be used to assess retrieval strategies aimed for result diversification. In the system-centred evaluation defined by controlled laboratory-based settings, a system uses a set of ambiguous or underspecified queries to *automatically* perform retrieval. Documents returned for each query are then evaluated against the known relevance judgements with respect to sub-topics (instead of documents) in terms of various performance measures, such as sub-topic recall, α -nDCG, Intent-Aware expected reciprocal rank (*ERR-IA*), etc.

Within this context, α -nDCG is one of the main official performance measures used at TREC 2009, 2010, and 2011. The measure is characterised by a parameter α , which

¹ Quote extracted from the TREC 2009–11 Web Diversity track guidelines [Clarke et al., 2009a, 2010].

²See Appendix D.

determines the balance between rewarding relevancy and intent coverage. In the TREC Web Diversity tracks, α is set to an *arbitrary* value of 0.5. The same setting is used by numerous studies, employing α -nDCG to evaluate and tune diversification methods, see for example [Sakai and Song, 2011; Santos et al., 2011a]. However, the effect of the value of α specified by an arbitrary setting has yet been thoroughly investigated. We discover that common settings of α , i.e. $\alpha = 0.5$, may prevent the measure from behaving as desired when evaluating result diversification in specific circumstances. This is because it excessively penalises systems that cover many sub-topics while it rewards those that redundantly cover only a few sub-topics. We highlight that this issue is crucial because it affects obtained rank systems at the very top, and also because if α -nDCG is used for learning-to-rank, it will produce a document ranking that is not relevant to user preferences in diversity task.

7.3 User Models within the Diversity Task

In order to examine the intuitiveness of α -nDCG, we formalise user models for an evaluation framework of sub-topic retrieval. Our aim is to compare the effectiveness measured by α -nDCG with user models.

Here, we consider the requirements of the task of sub-topic retrieval, as defined by the TREC 2009-11 Web Diversity track guidelines. This task requires “a ranked list of pages that together provide a complete coverage for a query, while avoiding excessive redundancy in the result list”. Given these requirements, a specific user model for assessing system performance is somewhat yet unclear. In particular:

- Should a system be considered more effective than another when it ranks many documents that are relevant to a single query-intent, although the alternative system retrieves fewer relevant documents, but with a larger query-intents coverage?
- Should a system be deemed more effective than another when it presents little redundancy despite a smaller number of relevant documents?

Throughout our investigation on diversity measures, we make a number of working assumptions regarding the evaluation framework and specifically considering the user information seeking behaviours. To this aim, we analyse two different user models

Table 7.1: A document/sub-topic matrix representing the relevance judgements of a query q made on four documents d_1, d_2, d_3, d_4 with respect to four sub-topics s_1, s_2, s_3, s_4 . A full dot (i.e. •) in a cell (i, j) of the matrix represents the case where a document d_i has been found relevant to sub-topic s_j .

		Sub-topics			
		s_1	s_2	s_3	s_4
Documents	d_1	•			
	d_2		•		
	d_3			•	
	d_4				•

that stem from different assumptions. Nevertheless, we show that these user models lead to the same requirements put forward by the evaluation framework of the diversity task. The main point is that rankings that cover an entire array of sub-topics should be preferred over rankings that only partially address some sub-topics. Two user models associated with the diversity task are described as follows.

7.3.1 User Model 1 – a set of users and a single document

In line with the assumption of TREC Web Diversity track, we assume a user model for a *set* of users who submit the same query, but with different query-intents. We refer this user model to the category of extrinsic diversity as discussed in 1.1.2. For example, the ambiguous query “window” can be interpreted differently depending on the user’s intent. A computer science student who issues this query is likely interested in the topic “Windows operating system”, whereas a professional glass artist is quite probably not to be interested in such a topic, and would judge documents about the operating system as non-relevant. The artist may consider relevant documents that discuss the topic of “stained glass window”. In this user model, it is assumed that users enter an ambiguous query that can have multiple interpretations, but each users is only interested in one of these. This user model has underpinned the work of [Maron and Kuhns \[1960\]](#) and [Stirling \[1977\]](#), where the relevance of documents is considered as a relationship between a set of users and a single document. In the following we provide an example scenario related to ranking documents under this user model.

Now, suppose an unknown user u (i.e. the system does not know which intents the user u has) issues a query q and the system retrieves four documents for the query. For

Table 7.2: Relevance judgements based on *User Model 1*. A document/user matrix representing the relevance judgements made on four documents (i.e. d_1, \dots, d_4) by ten users (i.e. u_1, \dots, u_{10}), issuing the same query q . Each user is interested in only one sub-topic, where their sub-topic relevance judgements correspond to the matrix presented in Table 7.1. A full dot (i.e. \bullet) in a cell (i, j) represents the case where document d_i has been found relevant by user u_j .

	Users										$P(R d, q)$
	u_1	u_2	u_3	u_4	u_5	u_6	u_7	u_8	u_9	u_{10}	
Documents	d_1	\bullet				\bullet		\bullet	\bullet		4/10
	d_2		\bullet				\bullet				2/10
	d_3				\bullet						1/10
	d_4		\bullet	\bullet						\bullet	3/10

each document, it is assumed that the system has previously collected relevance judgements with respect to each sub-topic (intent) s associated with a query q . In Table 7.1, we provide an example of those four documents and their relevance judgements, where a document d_1 is judged relevant to only a sub-topic s_1 , d_2 to s_2 , d_3 to s_3 , and d_4 to s_4 .

Next, ten users enter the same ambiguous query and the system returns four documents in response to such a query. Table 7.2 shows an example of this situation, where each document (a row in the table) has been judged relevant by each user (a column in the table, as denoted by u_1, \dots, u_{10}). Each individual user has a particular information need associated with a single sub-topic, and is not interested in the others. Correspondingly, relevance judgements of documents can be characterised with respect to each single user, or to a set of users who share a common feature or have a similar information need. For example, a document d_2 is judged relevant by users u_3 and u_7 , but not relevant by the remaining users. This is due to the fact that users u_3 and u_7 are interested in only the sub-topic s_2 contained in a document d_2 (see Table 7.1).

In which order should the retrieval system return documents to an unknown user u who issues a query q ? If the user u was *known* and, for instance, was user u_3 , the *ideal* retrieval strategy would be to retrieve, with respect to the query q , *all and only* documents that have been previously judged relevant to the sub-topic s_2 . Nonetheless, if user u_1 arrived and issued the same query q with intention to sub-topic s_1 , a system that used the *same* retrieval strategy would be considered ineffective. This is because it did not retrieve any relevant documents for the user u_1 . On the contrary, if a system

used the different retrieval strategy *diversifying* results to cover all four sub-topics, users u_1 and u_3 would consider this system effective enough to provide them with some relevant documents.

Returning to the previously posed question, how should the example documents in Table 7.2 be ranked in a result list? In ad-hoc retrieval, Stirling [1977]¹ suggested that the optimal retrieval strategy is to rank documents in order of decreasing probability of relevance, where the probability of relevance is computed according to the mean of relevance over the set of users who have judged the same documents relevant (the right-most column in Table 7.2). For instance, the document d_1 has been deemed relevant by four out of ten users, and thus is likely to be relevant to the user u with probability equal to 0.4. As a result, the optimal ranking for ad-hoc retrieval should be d_1 , followed by d_4 , d_2 , and d_3 . However, in sub-topic retrieval, if there was another document d_5 that is similar to d_1 (i.e. covers a sub-topic s_1), then at what position should the document d_5 be ranked in the list? Should it be ranked next to the document d_1 since it would obtain the same probability of relevance as document d_1 ? An ideal retrieval strategy for sub-topic retrieval is to return the document d_5 after documents that, when taken together, cover all sub-topics (e.g., d_1, \dots, d_4).

In sub-topic retrieval, Agrawal et al. [2009] proposed a family of *Intent Aware* (IA) metrics that comprise the probabilities of intents (sub-topic) or the likelihood that a query q is interpreted to a sub-topic s . The IA metrics assume that every user has a single intent. The probabilities of intents indicate the importances of each sub-topic and therefore how sub-topics (i.e. documents that cover such sub-topics) should be ranked. One way to determine the probability of intent is to derive from the probability of relevance (cf. [Stirling, 1977]) or how popular users are interested in a particular sub-topic. Regardless of the intent popularity, retrieval systems should respond to such a query of different users in a fair fashion. That is, every user should have an *equal* opportunity to find relevant information based on their own interest.

¹In his work, Stirling refers to this ranking criterion as the Probability Ranking Rule. Apart from the different interpretation of probability of relevance, the Probability Ranking Rule resembles the ranking criterion of the Probability Ranking Principle.

Table 7.3: Relevance judgements based on *User Model 2*. A document/user matrix representing the relevance judgements made on four documents (i.e. d_1, \dots, d_4) by users (i.e. u_{11}, u_{12}), issuing the same query q . Each user is interested in more than one sub-topics and their sub-topic relevance judgements correspond to the matrix presented in Table 7.1. A full dot (i.e. \bullet) in a cell (i, j) represents the case where document d_i has been found relevant to sub-topic s_j .

		Users			
		u_{11}			
Documents	d_1	\bullet			
	d_2		\bullet		
	d_3			\bullet	
	d_4				\bullet
		s_1	s_2	s_3	s_4
		Sub-topics			

		Users			
		u_{12}			
Documents	d_1	\bullet			
	d_2		\bullet		
	d_3				
	d_4				\bullet
		s_1	s_2	s_3	s_4
		Sub-topics			

7.3.2 User Model 2 – a single user and a set of documents

In the formulation and analysis of rank-based evaluation measures (e.g. $nDCG$ [Järvelin and Kekäläinen, 2002], RBP [Moffat and Zobel, 2008], ERR [Chapelle et al., 2009]), a user model is assumed for a *single* user who examines a *set* of ranked documents in a linear fashion and judges all documents relevant to their information need¹. The purpose of this user model is to assume the utility of retrieval systems by reflecting user's interactions with documents retrieved by the systems. Furthermore, a relationship between a single user and a set documents has been adopted to model the (probability of) document relevance in the development of several IR models (e.g. Probability Ranking Principle [Robertson, 1977; Robertson and Belkin, 1978; Robertson and Spärck-Jones, 1976] and Language Model [Hiemstra, 2011; Ponte and Croft, 1998]). Furthermore, this user model is in accordance with the category of intrinsic diversity (see section 1.1.1), in which diversity is considered a property of information need. That is, the user of intrinsic diversity requires a *set* of different results, taken together, to fulfil their *single* well-defined information need.

¹Some measures assume documents are independently judged (e.g. $nDCG$) whereas others do not (e.g. RBP and ERR).

Consider the example given earlier for a user model 1: a computer science student, who poses the query “windows” and is currently doing a report about key technologies behind new operating systems, may consider many relevant documents. These documents may cover several coherent topics of Windows 8, such as touch-centric user interface, cross-platform support (x86 and ARM processors), improved search functionality, etc. On the contrary, he may consider documents that discuss stained glass or window frame (i.e. not computer related) non-relevant.

In the context of sub-topic retrieval, we assume that there is a *single* user, who is most interested in a complete coverage of sub-topics of interest and is reluctant to see documents containing redundant (despite relevant) sub-topics that have been seen before. However, he still prefers documents covering *redundant relevant* sub-topics to *non-relevant* documents because the former documents are still considered more useful than the latter. In addition, he could expect to discover all relevant sub-topics after examining documents up to a cut-off rank r ¹. Therefore, the user preference of this user model is:

$$\text{“sub-topic coverage”} > \text{“redundant relevant”} > \text{“non-relevant”}$$

In other words, at document cut-off r , this user deems a system that retrieves fewer relevant documents but covers all sub-topics, more effective than a system that retrieves only relevant but redundant documents without providing a broad coverage of the query. Furthermore, given the same level of sub-topic coverage at a specific rank r , this type of user prefers systems that retrieve more relevant documents, to systems that retrieve fewer relevant documents.

In Table 7.3, we provide an example of user model 2, where a user u_{11} intends to find all four relevant sub-topics (Left), and a user u_{12} is interested in finding only three out of the four sub-topics (Right). A user u_{11} can be thought as an exploratory searcher² who want to find all possible aspects of a search topic. A user u_{12} can be

¹This has an implication on the importance of a ranking position at which retrieval systems are evaluated in this context. That is, if we assess retrieval systems at rank $r = 10$, it implies that within 10 examined documents, a user prefers to see all relevant sub-topics rather than to get all relevant documents that cover a few sub-topics.

²A user in exploratory search as assumed in the TREC interactive track [Over, 2001].

considered as a user who poses a faceted or underspecified query¹ and is interested in one sub-topic, but may still be interested in others as well (i.e. one or more sub-topics but not necessary all). Note that we consider the user u_{12} as a subset or specific case of this user model.

If the user in the user model 2 was *unknown*, a system that employed a typical retrieval strategy might retrieve documents covering only some sub-topics. Such a user would have considered the system *partially* effective because it returns only some of the relevant sub-topics for him. In comparison with the user model 1, the user of user model 2 could expect to find *all* or *more than one* relevant sub-topics associated with his multi-aspect information need (e.g. s_1, \dots, s_4) when examining r documents. r is a ranking position at which retrieval systems are evaluated, and we assume the user's expectation in finding all desired sub-topics.

7.3.3 Recap of the Two User Models

We had described two users models related to sub-topic retrieval. Although these two user models are fundamentally different, they both can be applied for ranking and evaluation approaches with little adaptations. In the rest of the thesis, we will adopt the second one which is the common user model adopted in current IR research. We then can validate evaluation measures that account for novelty and diversity in search results.

Recall that the type of user u is unknown and data for guessing u 's user model are unavailable. Therefore, a user u is equally likely to be any of the users u_1, \dots, u_{12} , or in turn any of the user models. Within this situation, a system would have to take a risk on retrieving documents that belong to all relevant sub-topics, which the user u may be interested in.

7.4 Evaluation Measures

Evaluation measures for sub-topic retrieval attempt to quantify the degree to which retrieval results address the breadth of possible intents of information needs underlying a query. Many proposed measures were built with the common assumption that defines

¹Another type of information needs, which is expressed in the form of faceted queries as defined in TREC Web Diversity track.

query-intents as a set of individual sub-topics. A standard approach to diversity evaluation involves a collection of documents and a set of information needs expressed by broad or ambiguous queries (topics). Each query is composed of a representative set of sub-topics, each of these relates to a different query-intent. Documents are judged to be relevant with respect to each sub-topic. A set of queries serves, together with the corpus, as input for retrieval system and its output (i.e. ranked documents returned for each query) is evaluated against the sub-topic relevance judgements. While in some measures diversity is *explicitly* evaluated in terms of the *coverage* of sub-topics in a result list, in other measures diversity is *implicitly* evaluated through the *redundancy* of sub-topics. In this section, we describe in details the two main families of evaluation measures of sub-topic retrieval, i.e. measures based on *i*) diversity and *ii*) redundancy.

7.4.1 Diversity-Based Measures

In this section, we describe two evaluation measures based upon diversity. These measures *explicitly* evaluate diversity with respect to the coverage of sub-topics of a query.

7.4.1.1 Sub-topic Recall

Zhai et al. [2003] generalised the traditional relevance-based measures of precision and recall. They simplify different aspects of relevant information as “sub-topics” of a general topic. Particularly, *Sub-topic recall* (*s-recall*) is defined in analogy to traditional recall in order to measure the effectiveness of IR systems in terms of the fraction of sub-topics that are covered by the retrieved documents. Suppose that a query q is composed of $|S|$ sub-topics $S = \{s_1, \dots, s_{|S|}\}$, and that a document can either be relevant or non-relevant to each sub-topic. Let d_i denote the document retrieved at rank i , and $sub-topic(d_i)$ be the set of sub-topics to which d_i is relevant. Then, *s-recall* at cut-off r can be defined as the proportion of sub-topics covered by the top r documents with respect to the total number of sub-topics associated with q , i.e.:

$$s-recall@r = \frac{|\cup_{i=1}^r sub-topic(d_i)|}{|S|} \quad (7.1)$$

S-recall explicitly evaluates the diversity of relevant information in terms of sub-topic coverage in a document ranking. Therefore, the greater s-recall is, the higher the number of *different* sub-topics covered in a ranking. However, when used on its own, s-recall is a rather coarse metric because it is affected by three major drawbacks.

- 1) Similar to traditional recall, s-recall is a non-position based metric and thus does not account for the positions at which relevant documents are retrieved. In other words, at a cut-off rank r , s-recall does not distinguish the benefit of retrieving a relevant document at rank $r - 1$ or at rank $r - 2$. In fact users prefer relevant information to be retrieved as early as possible, and so the relevant document at rank $r - 2$ should be considered better than that at $r - 1$.
- 2) Once a sub-topic has been covered, s-recall does not distinguish between subsequent retrievals of (documents covering) the same relevant sub-topic. Although retrieving a redundant *relevant* sub-topic may be undesirable for users, it is still considered to be more useful than retrieving a *non-relevant* sub-topic.
- 3) The second issue generalises into the third drawback of s-recall: once all sub-topics are covered, s-recall does not further distinguish between retrieving relevant or non-relevant documents. In fact s-recall is only able to address the sub-topic diversity upto the position at which all relevant sub-topics are retrieved. After complete sub-topic coverage is achieved, s-recall always equals 1 and thus it cannot identify how well retrieval systems diversify search results.

7.4.1.2 Sub-topic Mean Reciprocal Rank

Similar to a traditional *mean reciprocal rank* (*mrr*) [Voorhees, 1999], *sub-topic mean reciprocal rank* (*s-mrr*) is defined as the inverse of the rank at which a specific percentage of sub-topic coverage is achieved (e.g. 25%, 50%, 75%, 100%, etc.) [Chen and Karger, 2006; Wang and Zhu, 2009; Zuccon and Azzopardi, 2010]. Let Q be a sample set of queries $q_1, \dots, q_{|Q|}$ and p be a percentage of sub-topics covered by documents at a cut-off r_q given a query q . $rank_q$ is the first rank r_q at which s-recall achieves p , defined as:

$$rank_q = \min(r_q : s-recall@r_q \times 100 > p\%)$$

Then, we can define s-mrr as the average of the reciprocal ranks of results for a sample of queries Q :

$$s-mrr@p\% = \frac{1}{|Q|} \sum_{q=1}^{|Q|} \frac{1}{rank_q} \quad (7.2)$$

S-mrr measure has the benefit of measuring systems that attempt to cover all sub-topic as early positions as possible. However, s-mrr only partially tackles the first drawback of s-recall. Although s-mrr is computed based on a document position, it takes into account *only* the first position at which the specified s-recall is achieved. Positions of other documents retrieved before or after that position are ignored and not included in the computation. For example, consider two document rankings for a topic that contains two relevant sub-topics. S-mrr measures that these two rankings cover all relevant sub-topics at position $r = 3$. Nevertheless, one ranking covers the first sub-topic at rank $r = 1$, and the other ranking covers the first sub-topic at rank $r = 2$. In fact, the former should be considered better than the other, but s-mrr cannot differentiate the effectiveness of these two rankings.

Note that both s-recall and s-mrr assume binary relevance judgements, thus not accounting for graded relevance. In practice, they cannot then differentiate between highly and marginally relevant documents.

7.4.2 Redundancy-Based Measures

Traditional IR measures that stem from models of user-browsing behaviour, such as $nDCG$ [Järvelin and Kekäläinen, 2002], RBP [Moffat and Zobel, 2008], ERR [Chapelle et al., 2009], have been extended to the evaluation context of novelty and diversity retrieval (see $\alpha-nDCG$ [Clarke et al., 2008], $NRBP$ [Clarke et al., 2009b], and $ERR-IA$ [Chapelle et al., 2009]). All these measures are characterised by a *similar* gain function that models the documents' utility. Moreover, each measure is distinguishable because of the different *discount* functions that progressively reduces document

utility with respect to rank positions or user's effort [Carterette, 2011]. In some of these measures, the discount functions (or *additional* discount functions) are viewed as modelling user's effort to endure examining *redundant* documents. Through the functions penalising redundancy, diversity is *implicitly* evaluated, with an expectation to reward a system that retrieves documents containing less redundant information.

7.4.2.1 Novelty Biased Discounted Cumulative Gain

Novelty Biased Discounted Cumulative Gain (α -nDCG) has been developed as a modification of *normalised discounted cumulative gain* (nDCG) so as to accommodate for the evaluation of novelty and diversity [Clarke et al., 2008]. The formalisation of the measure revolves around the concept of nuggets, which represent the information needs associated with query-intents. A document is considered relevant if it contains any relevant information. In other words, a particular document is relevant if it contains at least one nugget that is also contained in the users information need. Each nugget is assumed to be independent and the probability that a document contains a nugget determines the graded relevance of a nugget. However, following the judgements expressed by TREC human assessors, Clarke et al. [2008] further assumed a binary decision regarding each nugget: i.e. Is the nugget contained in the document or not? (see Section 4.1 in [Clarke et al., 2008]). In line with this assumption, we shall use binary judgements when deriving an approach for setting alpha.

The formalisation of α -nDCG proceeds as follows. Consider a query q with a total of $|S| > 1$ sub-topics or query-intents of a user. Let n_k be a nugget and $|N|$ be the total number of nuggets contained in documents and associated with a query q . Sub-topic s is a set of nuggets, represented in the form $s_1, \dots, s_{|S|} \supseteq \{n_1, \dots, n_{|N|}\}$, where $|S| \leq |N|$. That is, for instance, there might be $|S| = 2$ and $|N| = 3$. It may be that $s_1 = \{n_1, n_2\}$ and $s_2 = \{n_3\}$. Nevertheless, in practice (e.g. in TREC evaluation [Clarke et al., 2009a, 2010]), sub-topics and nuggets are commonly used interchangeably on the assumption that sub-topics are consistent with nuggets. This assumption leads to the correspondences of the two definitions, i.e. $s_1 = \{n_1\}$, $s_2 = \{n_2\}, \dots, s_{|S|} = \{n_{|N|}\}$ and $|S| = |N|$.

Let $J(d_r, s) = 1$ if a document d_r at rank r is relevant to a sub-topic s and 0 otherwise. Then, a *duplication* measure $D_{s,r-1}$ can be defined as:

$$D_{s,r-1} = \begin{cases} \sum_{i=1}^{r-1} J(d_i, s) & \text{if } r > 1 \\ 0 & \text{if } r = 1 \end{cases}$$

The role of $D_{s,r-1}$ is to monitor the degree of duplicate or redundant information within the documents ranked above rank r , given a sub-topic s . In other words, when $r > 1$, $D_{s,r-1}$ is the number of times a sub-topic s appeared in documents ranked within the top $r - 1$. The measure has the role of quantifying the benefit of a document in a ranking. We call this the *novelty-biased gain*, $NG(q, r)$:

$$NG(q, r) = \sum_{s=1}^{|S|} J(d_r, s)(1 - \alpha)^{D_{s,r-1}} \quad (7.3)$$

where the parameter $0 < \alpha < 1$ ¹ represents the probability of assessor error, i.e. the likelihood that the assessor incorrectly judges that a document d contains a relevant sub-topic s . In practice, the parameter is used to define the probability of user's intolerance to a redundant relevant sub-topic. Therefore, α manipulates the amount of penalisation to assign to a document carrying redundant information. The higher the value of α is, the greater the discount applied to documents containing redundant sub-topics. To account for the late retrieval of documents containing relevant sub-topics, [Clarke et al. \[2008\]](#) compute a gain at rank r based on $NG(q, r)$, instead of the traditional gain that directly reflects the graded relevance value of a document [[Järvelin and Kekäläinen, 2002](#)]. The modified gain is further discounted by dividing with respect to a function of the rank position, called *normalised utility* [[Sakai and Robertson, 2008](#)]. The discount function, e.g. $\log_2(1 + r)$, reflects the decay weight utility of relevant sub-topics that are less likely to be examined by users. The gain is then progressively cumulated, obtaining the discounted cumulative gain, $DCNG(r)$:

$$DCNG(r) = \sum_{i=1}^r \frac{NG(q, i)}{\log_2(1 + i)} \quad (7.4)$$

¹We excluded the value $\alpha = 0$, since α -nDCG's function would be equivalent to that of nDCG. $\alpha = 1$ was also not considered since the gain $NG(q, r)$ would be equal to zero and unable to be identified.

The discounted cumulative gain at rank r is finally normalised by that of the ideal ranking r^* or $DCNG(r^*)$, which maximises $DCNG(r)$. α -nDCG can be then defined as:

$$\alpha - nDCG = \frac{DCNG(r)}{DCNG(r^*)} \quad (7.5)$$

Through the duplication measure $D_{s,r-1}$, α -nDCG *explicitly* accounts for the redundancy of *relevant* sub-topics that have been retrieved previously. It discounts the gain obtained from relevant sub-topics based on the degree of redundancy of such sub-topics covered by a document at position $r - 1$. While redundancy is addressed explicitly in the measure, [Clarke et al. \[2008\]](#) suggested that *diversity* is included by accumulating the gain of relevant sub-topics (nuggets) present in documents. Nevertheless, it can be argued that diversity of sub-topics is simply ignored and dominated by the relevance to sub-topics. In fact, α -nDCG has been built based upon a series of assumptions, that we analyse below.

Assumption 1: α -nDCG assumes that each sub-topic is independently judged and thus the relevance of a sub-topic (e.g. s_1) does not depend on that of other sub-topics (e.g. $s_2, \dots, s_{|S|}$). Similarly, the same approach is used when computing the gain for each sub-topic. The gain that α -nDCG assigns to a document covering a sub-topic is only based on how many times the sub-topic has been already covered by previously ranked documents. In particular:

- The more times a sub-topic is covered, the higher the gain of new documents covering the same sub-topics is discounted.
- The gain, as well as the amount of discount, is computed independently for each sub-topic.

This may produce a situation where a document covering some redundant¹ sub-topics (e.g. 2) is assigned a higher gain than a document covering a missing² sub-topic.

¹i.e. sub-topics that have been already covered by previously ranked documents.

²i.e. a relevant sub-topic that has not been covered by any document retrieved so far.

This may occur, although the gain of the former document is discounted and the one of the latter is not. Therefore, the amount of discount assigned to redundant sub-topics is too small to sufficiently decrease the gain they provide. On the other hand, it may happen that a document containing a missing sub-topic is not rewarded enough for covering the missing sub-topic.

As a result, α -nDCG evaluates systems that return many relevant sub-topics but little diversity as being superior to those that attempt to cover all sub-topics for a given query. Due to this issue, Sakai and Song [2011] also argued that α -nDCG is counter-intuitive; however, they did not identify in which circumstances this is the case. While we agree with their claims, in this thesis we take a step further in the analysis of the measure. We in fact identify the circumstances where α -nDCG is counter-intuitive, and we explain why this is so. Then we propose a solution to improve the intuitiveness of α -nDCG, without resorting to developing a new measure, as opposed to [Sakai and Song, 2011]. We will discuss this issue in detail in the following sections, where we illustrate the problematic circumstances and propose an approach to solve the problem.

Assumption 2: The original formulation of α -nDCG assumes a uniform probability for all different query-intents or sub-topics (See Section 4.2 in [Clarke et al., 2008]). Agrawal et al. [2009] proposed a family of *Intent-Aware* (IA) metrics, accommodating the probabilities of intents $P(i|q)$, where i is an intent of a query q . They generalise traditional metrics such as MAP, nDCG, etc., by factorizing in their formulation the intent probabilities, which represent the likelihood that a user is searching for specific intents given the issued query. By including the intent probabilities, a measure is supposed to prefer a system, retrieving a popular sub-topic at early ranks since a user is highly likely to search for such a sub-topic compared to other less popular topics. Clarke et al. [2011] later suggested that their α -nDCG can be extended to incorporate the intent probabilities. However, throughout this thesis we restrict our investigation in line with the TREC 2009–11 Web Diversity track by considering a *uniform* intent-probability distribution, i.e. the user is equally interested in each one of the identified query-intents.

7.5 Analysis of α -nDCG

α -nDCG is characterised by a parameter α , which sets the balance between the reward of relevant information and the detriment of redundant information within the retrieved documents. In common IR evaluation contexts, such as in the TREC 2009–11 Web Diversity tracks, α is set to an arbitrary value of 0.5. A little study on how α affect rankings was investigated by [Clarke et al. \[2008\]](#). They varied the value of α equally over all queries in the range $[0,1)$ with a step of 0.25. They only studied the effect of α on computing the α -nDCG score on a re-ordering of the document ranking, called a *reversed ideal gain vector*¹. However, no comprehensive studies have been conducted to investigate the impact of α (i.e. $\alpha = 0.5$) on assessing document rankings, and in particular to understand how to appropriately set α that depends on the sub-topics of a query.

In this section, we show that arbitrarily setting α to 0.5 may be a misleading practice, as the measure might turn to be counter-intuitive and not behave as anticipated by the evaluation guidelines. We show in fact that in some circumstances and adopting the common settings ($\alpha = 0.5$), α -nDCG tends to reward systems that retrieve redundant relevant documents which offer only a partial sub-topics coverage. It instead penalises systems that successfully provide a complete coverage of all the relevant query-intents. We uncover this issue by showing an example scenario and emphasise that the issue is crucial as it highly influences systems ranked on top, i.e. the best performing systems.

Table 7.4 shows five documents relevant to (some of) four sub-topics of query 26 belonging to the TREC 2009 Web Diversity Track. The query topic is “lower heart rate”. This topic is accompanied by four different sub-topics, i.e.

26.1: “What causes the heart to beat faster or slower?”,

26.2: “What is a normal heart rate when a person is resting?”,

26.3: “How can I lower my heart rate?”, and

26.4: “Is a higher heart rate related to high blood pressure or cholesterol?”.

¹The vector of a document ranking, constructed by using a greedy algorithm to minimise the α -nDCG of relevant documents

Documents that are considered relevant have to contain answer(s) regarding at least one of the four sub-topics. For example, documents *a* and *c* cover three relevant sub-topics each (i.e. {26.1, 26.3, 26.4}), document *d* covers two sub-topics (i.e. {26.3, 26.4}), and so on. Notice that document *b* contains only one relevant sub-topic, i.e. 26.2, which is not covered by the other four documents in the example.

Table 7.4: Five documents relevant to the sub-topics of query 26, “lower heart rate”, from the TREC 2009 Web Diversity Track.

Document ID	Sub-topic				Total
	26.1	26.2	26.3	26.4	
<i>a.</i> “clueweb09-en0001-55-27315”	1	-	1	1	3
<i>b.</i> “clueweb09-en0004-47-03622”	-	1	-	-	1
<i>c.</i> “clueweb09-en0001-69-19695”	1	-	1	1	3
<i>d.</i> “clueweb09-en0003-94-18489”	-	-	1	1	2
<i>e.</i> “clueweb09-en0000-31-13205”	-	-	-	-	0

To illustrate the situation where α -nDCG behaves counter-intuitively, we consider three imaginary system rankings (*A*, *B*, *C*), where the top three documents are ranked differently. In Table 7.5, the first column shows the rank position, (*r*), followed by document id, (*doc*), and the gain, *g*(*r*), with respect to sub-topic relevance. The other columns report respectively the novelty-biased gain, *ng*(*r*), discounted novelty-biased gain, *dng*(*r*), the discounted cumulative novelty-biased gain, *dcng*(*r*), its normalised gain, α -ndcg(*r*) when $\alpha=0.5$, and finally the sub-topic recall, *s-r*(*r*).

An ideal ranking for α -nDCG is a document order that provides the maximum *dcng*(*r**) score at a position *r* in a ranking. An example of this ranking is shown in the last row of the table.

Note that α -nDCG has worst-case NP-hard computation time to obtain the maximum value of *dcng*(*r**) [Agrawal et al., 2009; Carterette, 2009]. Therefore, we commonly resorted to a greedy algorithm¹ that produces a local optimum at each point in a ranking. Although the greedy algorithm is not optimal and may lead to over-rate a bad system, Carterette [2009] estimated that only 7% of queries are sub-optimal in the TREC 2009 Web Diversity track corpus. Therefore, since for most of the queries

¹i.e. maximising the gain (i.e. *dng*(*r*)) at each rank position without revising the choice made at previous rank.

Table 7.5: Corresponding evaluations of three imaginary system rankings for query 26 using α -nDCG, when $\alpha=0.5$ and an ideal ranking, of which $dcng(r)$ are used for normalisation

		r	doc	g(r)	ng(r)	dng(r)	dcng(r)	α -ndcg(r)	s-r(r)
system	A	1	a	3	<u>3.00</u>	3.00	3.00	1.00	0.75
		2	c	3	<u>1.50</u>	0.95	3.95	1.00	0.75
		3	e	0	<u>0.00</u>	0.00	3.95	0.89	0.75
	B	1	a	3	<u>3.00</u>	3.00	3.00	1.00	0.75
		2	d	2	<u>1.00</u>	0.63	3.63	0.92	0.75
		3	e	0	<u>0.00</u>	0.00	3.63	0.82	0.75
	C	1	a	3	<u>3.00</u>	3.00	3.00	1.00	0.75
		2	b	1	<u>1.00</u>	0.63	3.63	0.92	1.00
		3	e	0	<u>0.00</u>	0.00	3.63	0.82	1.00
<i>ideal</i>		1	a	3	<u>3.00</u>	3.00	3.00	1.00	0.75
<i>ranking</i>		2	c	3	<u>1.50</u>	0.95	3.95	1.00	0.75
$\alpha = 0.5$		3	b	1	<u>1.00</u>	0.50	4.45	1.00	1.00

this problem does not occur, we employ the greedy algorithm for all the queries as a trade-off between accuracy and computational complexity.

A user's ideal ranking may differ from that of α -nDCG. The ideal ranking of a user is a specific order in which a user expects a system to retrieve documents so as to satisfy his information needs associated with different sub-topics. To obtain this ideal ranking, click log analysis has been suggested to match a document ordering with observed click behaviour [Chapelle et al., 2009; Yilmaz et al., 2010; Zhang et al., 2010]. Otherwise, by employing crowdsourcing platforms (e.g. Amazon Mechanical Turk (AMT)¹ or CrowdFlower²), researchers can ask users (or workers) which document lists they prefer in order to find a user's ideal ranking. For example, Sanderson et al. [2010] used AMT to gather *user preferences* about rankings of documents and then studied the correlation between the collected preferences and the evaluations of the same rankings obtained using different IR metrics.

Another method to generate a user's ideal ranking is through user modelling. We employed user models in Section 7.3 to hypothesise a particular ranking a user prefers.

¹<http://www.mturk.com/>

²<http://crowdflower.com/>

Consider again the example of Table 7.4, consisting of five documents. An ideal document ranking for users, as assumed by our user models, is $a-b-c-d-e$ or $c-b-a-d-e$. This is because the main goal is to cover all sub-topics in the most early positions. Then, due to carrying more useful information to users, the secondary goal is to reward documents containing more relevant sub-topics higher than those documents containing fewer relevant sub-topics. Therefore, documents a or c , covering the most number of relevant sub-topics (i.e. 3), should be ranked in the first position. Document b should be retrieved next because it covers a missing sub-topic 26.2 that is not covered by any other document. The remainder positions are $c-d-e$ or $a-d-e$ since they are ranked according to the number of relevant sub-topics they contain.

Returning to the example of Table 7.5, while $a-b-c-d-e$ (or $c-b-a-d-e$) is an ideal ordering of the documents for *users*, setting α to 0.5 produces a maximal gain that prefers a *non* ideal document ranking $a-c-b-d-e$. Note that the $\text{dcng}(r^*)$ of the α -nDCG's ideal ranking (highlighted in italic) is in turn used for normalising those of imaginary system rankings to obtain the final $\alpha\text{-ndcg}(r)$. Therefore, if systems are evaluated according to $\alpha\text{-nDCG}@3$ with $\alpha=0.5$, the following system rankings are obtained: (A, B, C) or (A, C, B) . This is counter-intuitive and does not reflect the user preference towards the system. In fact system C obtains a lower $\alpha\text{-nDCG}$ than system A at both rank 2 and 3, although at rank 2 it covers the only missing sub-topic (26.2) achieving complete sub-topic coverage (i.e. $s\text{-}r(2)=1.0$) earlier than A . Here, $\alpha\text{-nDCG}$ with $\alpha=0.5$ rewards documents containing *novel* relevant sub-topics *less* than *redundant* sub-topics. As a consequence, a common $\alpha\text{-nDCG}$, developed based on user model 2, makes little sense in the context of web search, which applies best to user model 1. *Multiple* users issue intrinsically ambiguous queries and each user aims to find a *single* document relevant to his intent. Thus, sub-topic recall is better suited to the formal problem of sub-topic retrieval, i.e. measure a complete coverage of relevant sub-topics. However, the question arises: “How can we alleviate the issue of $\alpha\text{-nDCG}$ in order to reward more the documents containing novel relevant sub-topics and penalise more the documents containing redundant sub-topics?”

7.6 Deriving a Safe Threshold for α

In this section, we propose a solution to circumvent the issue that affects α -nDCG. The solution consists in setting the parameter α considering not only the evaluation preferences expressed by the user¹, but also the number of intents a query has.

We consider the user models of Section 7.3, and examine the case where the gain obtained by a system retrieving novel relevant sub-topics, say system X , is expected to be higher than the gain of a system retrieving only redundant sub-topics, say system Y . Now, let s^* be a novel relevant sub-topic (or the sub-topic with the smaller degree of redundancy), and s a redundant relevant sub-topic. We now consider the worst case scenario that may occur at a rank position r . This occurs when system X retrieves a document covering only a single *novel* relevant sub-topic, whereas system Y retrieves a document containing the remainder $|S|-1$ relevant but *redundant* sub-topics. In such situation, system X should obtain a higher α -nDCG than system Y because system X attempts to achieve a complete coverage of sub-topics. Thus, since we expect $NG_X(r) > NG_Y(r)$, we can rewrite this as:

$$J(d_r, s^*) \cdot (1 - \alpha)^{D_{s^*, r-1}} > \sum_s^{|S|-1} J(d_r, s) \cdot (1 - \alpha)^{D_{s, r-1}} \quad (7.6)$$

This inequality can be used to define boundaries on the value of parameter α so that the inequality is true, i.e. a system retrieving novel relevant sub-topics is awarded with a higher α -nDCG than a system retrieving redundant sub-topics. At this stage we make a simplifying assumption that is compatible with the relevance judgements that have been collected in the TREC Web Diversity track: we assume a binary decision schema regarding the relevance of documents to each sub-topic. That is, relevance is assumed to be a dichotomous quantity, and a document can be assessed as being relevant to a query-intent or not relevant. Therefore, equation (7.6) becomes:

$$(1 - \alpha)^{D_{s^*, r-1}} > \sum_s^{|S|-1} (1 - \alpha)^{D_{s, r-1}} \quad (7.7)$$

¹Encoded in a user model, as for example that of Section 7.3.

$D_{s,r-1}$ is the number of times a sub-topic s is covered by all documents till rank $r - 1$. We further assume that $D_{s,r-1}$ for all *redundant relevant* sub-topics $|S| - 1$ are equal. This means that $(1 - \alpha)^{D_{s,r-1}}$ of all redundant relevant sub-topics are also identical. With this assumption, equation (7.7) becomes:

$$\begin{aligned} (1 - \alpha)^{D_{s^*,r-1}} &> (|S| - 1) \cdot (1 - \alpha)^{D_{s,r-1}} \implies \\ \implies \frac{1}{(|S| - 1)} &> (1 - \alpha)^{D_{s,r-1} - D_{s^*,r-1}} \end{aligned} \quad (7.8)$$

Let $\beta = D_{s,r-1} - D_{s^*,r-1}$ be the difference in redundancy levels. This is, measuring a relative amount of novel information in documents, where redundant sub-topics have a higher degree of redundancy than novel sub-topics, i.e. $D_{s,r-1} > D_{s^*,r-1}$, and thus $\beta > 0$. When relevance assessments are binary, β is an integer. Thus, we can rewrite equation (7.8) as a system of two inequalities:

$$\begin{cases} \alpha > 1 - \left(\frac{1}{|S|-1}\right)^{1/\beta} & \text{if } (1 - \alpha) > 0 \\ \alpha < 1 + \left(\frac{1}{|S|-1}\right)^{1/\beta} & \text{if } (1 - \alpha) < 0 \end{cases} \quad (7.9)$$

The case when $(1 - \alpha) < 0 \implies \alpha > 1$ can be ignored because by definition $\alpha < 1$ meaning the case can never occur. By examining the base $(1 - \alpha) > 0 \implies \alpha < 1$, we can derive the safe threshold (st) for α :

$$\alpha > 1 - \left(\frac{1}{|S| - 1}\right)^{1/\beta} = st \quad (7.10)$$

Inequality (7.10) is the necessary and sufficient condition that has to be satisfied if α -nDCG has to be expected to reward systems retrieving novel relevant sub-topics more than systems retrieving redundant sub-topics. The threshold of equation (7.10) is a function of the number of intents associated with a query. Queries that differ for the number of sub-topics generate different values of the threshold. Therefore, if α is set

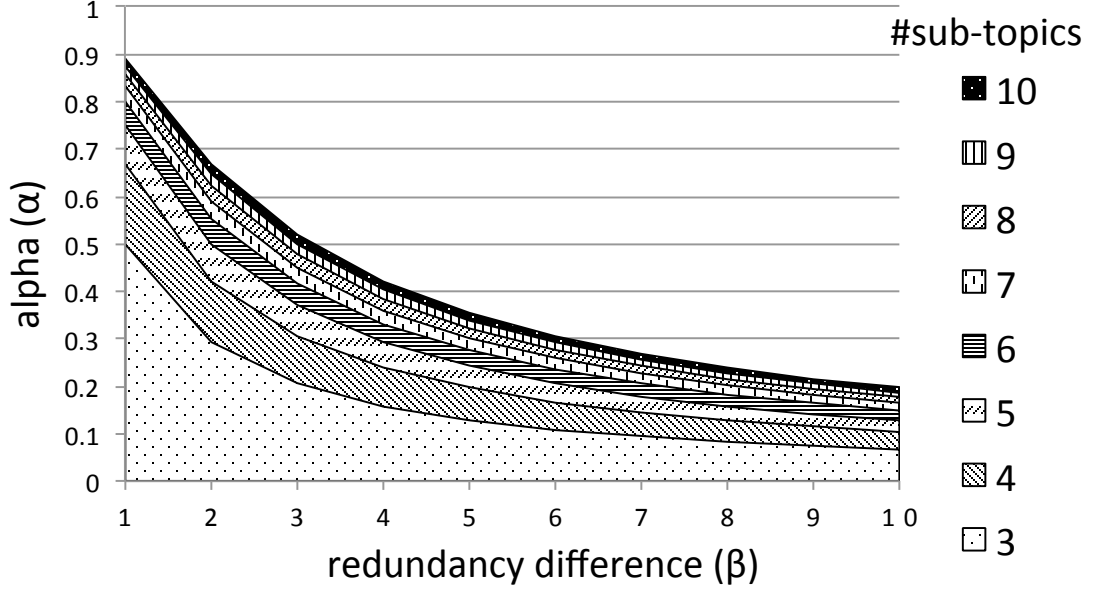


Figure 7.1: Values of the safe threshold for α .

to 0.5 regardless of the queries, α -nDCG may misjudge documents conveying novel (less redundant) information (recall an example shown in Table 7.5).

For queries containing 2 sub-topics, this problem does not occur, as $\alpha = 0.5$ is greater than the safe threshold. For queries with 3 or more sub-topics, values of α lower than the threshold violate the user preferences set by the user models derived from the TREC guidelines. This is because equation (7.10) suggests that α should be set greater than 0.5 when setting $|S| = 3$ and $\beta = 1$. Values of $\alpha \sim st + \epsilon$ (ϵ being a very small positive number) are the minimum values that satisfy the user models. Finally, for $\alpha \gg st$, increasing importance is given to diversity at the expense of relevance.

Figure 7.1 plots the safe threshold (st) on α according to equation (7.10) by varying circumstances, i.e. the number of sub-topics, the level of redundancy difference β , and the value of parameter α . The figure suggests that considering values of α below or equal to the threshold st (i.e. inside the highlighted areas) can lead to an unexpected behaviour of the measure. In an example of the query containing 10 sub-topics, Figure 7.1 suggests that α should be set greater than 0.89 when $\beta = 1$, 0.77 when $\beta = 2$, 0.52 when $\beta = 3$, and so on. Therefore, an upper bound of safe threshold that

can satisfy all necessary conditions is $\alpha > 0.89$. The threshold suggests that setting $\alpha = 0.5$ may lead α -nDCG to misjudge documents conveying novel information, as we discussed in Section 7.5. This is because α -nDCG with $\alpha = 0.5$ does not sufficiently discount the gain of redundant relevant sub-topics. This problem is crucial, in particular, when analysing *high quality*¹ ranking results @2, @3, etc., or when the redundancy difference of the rankings (β) at lower positions is small (e.g. $\beta = 1, 2$, or 3).

7.7 Examination of the Safe Threshold

Let us consider the same example scenario of Table 7.5 once again. Recall that if $\alpha = 0.5$ and systems are ranked according to α -nDCG@3, then the ordering (A, B, C) or (A, C, B) is found to be optimal despite not being as desired according to the user models in Section 7.3. Here, we revisit the behaviour of α -nDCG when the value of α is set according to the safe threshold (st) of inequality equation (7.10). For query 26, there are $|S| = 4$ sub-topics. By imposing $\beta = 1$ to define an upper bound of the threshold to avoid all cases in which $\alpha \leq st$ ², we obtain the value of $st = 0.67$ for the safe threshold. In empirical settings, we consider values of α with *double-precision* accuracy. Then the smallest ϵ is 0.01. We therefore obtain $\alpha = 0.68$ that is the smallest value for which the condition set by the safe threshold is satisfied.

In Table 7.6, we re-evaluate three imaginary systems using α -nDCG with $\alpha = 0.68$. The ideal ranking, obtained according to α -nDCG with α set by $\alpha = st + \epsilon$, is *a-b-c-d-e*. This ranking now corresponds to an ideal ranking of *users*, assumed by user preferences of user models in Section 7.3. In the last row of the table, we present the top three documents in the ideal ranking with their scores (i.e. $dcng(r^*)$, locally optimised by a greed algorithm) for normalisation. If α is set according to the safe threshold (e.g. $\alpha = st + 0.01$), a different retrieval performance is obtained.

By considering α -nDCG@3, we obtain a different system ranking, (C, A, B), which follows the TREC guidelines for the Web Diversity task. In fact, with the new setting of α , systems that provide complete sub-topic coverage are preferred to systems that

¹i.e. when relevant documents containing a large number of sub-topics are ranked within the early ranking positions.

² $\beta > 1$ always produce the safe thresholds lower than that of $\beta = 1$

Table 7.6: Corresponding evaluations of three imaginary system rankings for query 26 using α -nDCG, when $\alpha=0.68$ and an ideal ranking, of which dcng(r) are used for normalisation

		r	doc	g(r)	ng(r)	dng(r)	dcng(r)	α -ndcg(r)	s-r(r)
system	A	1	a	3	<u>3.00</u>	3.00	3.00	1.00	0.75
		2	c	3	<u>0.96</u>	0.61	3.61	0.99	0.75
		3	e	0	<u>0.00</u>	0.00	3.61	0.87	0.75
	B	1	a	3	<u>3.00</u>	3.00	3.00	1.00	0.75
		2	d	2	<u>0.64</u>	0.40	3.40	0.93	0.75
		3	e	0	<u>0.00</u>	0.00	3.40	0.82	0.75
	C	1	a	3	<u>3.00</u>	3.00	3.00	1.00	0.75
		2	b	1	<u>1.00</u>	0.63	3.63	1.00	1.00
		3	e	0	<u>0.00</u>	0.00	3.63	0.88	1.00
<i>ideal</i>		1	a	3	<u>3.00</u>	3.00	<i>3.00</i>	1.00	0.75
<i>ranking</i>		2	b	1	<u>1.00</u>	0.63	<i>3.63</i>	1.00	1.00
$\alpha = 0.68$		3	c	3	<u>0.96</u>	0.48	<i>4.11</i>	1.00	1.00

have less emphasis on the diversity of sub-topics. In these circumstances, α -nDCG with $\alpha = 0.68$ rewards the system C, which contains a missing sub-topic (i.e. 26.2), more than systems A and B, which at rank 3 cover only redundant sub-topics (i.e. 26.1, 26.3, and 26.4).

7.8 Summary

In this chapter, we have shown that arbitrarily setting the value of α (i.e. $\alpha = 0.5$) prevents α -nDCG from behaving as desired, i.e. reward systems that provide novel and diversified rankings. We proposed a theoretically sound approach which defines a formal threshold for the value of α on a query-basis. The key of our approach is to resolve the parameter setting of α -nDCG with respect to the number of sub-topics of each query. [Clarke et al. \[2008\]](#) say that α is a user parameter. However, with our derivation of the safe threshold, we show the value of α is conditioned on the number of query-intents or sub-topics. By doing so, α can be reported not only as a user dependent parameter, but also as a parameter depending upon sub-topics. In other words, the value of α increases as the ambiguity of queries increases (i.e. higher ambiguous queries can be interpreted into more different meanings). Therefore the

parameter α , which represents the probability of user's intolerance to redundant sub-topics, should also be specified according to the ambiguity of queries. Although α -nDCG is mainly devised on the basis of *redundancy*, we analytically show that our safe threshold allows the measure to evaluate the *diversity* of sub-topics in a document ranking.

Unlike s-recall, which neither accounts for redundant relevant sub-topics, nor addresses diversity after complete sub-topic coverage is achieved, α -nDCG with $\alpha > st$ can address the utility of rankings in terms of documents' rank positions, relevancy, redundancy, and importantly diversity. By examining the example scenarios in Table 7.6, we have shown that the use of the safe threshold allows α -nDCG to behave as desired, i.e. following the TREC guideline. Nevertheless, some key issues arise when our approach is used in the evaluation framework:

- How many times do circumstances similar to that of Table 7.5 occur when considering real data?
- Which systems are over-rated or under-rated when using α -nDCG with common settings (i.e. $\alpha = 0.5$)?
- How much do system rankings differ when using common settings and a safe threshold?
- How intuitive is the evaluation behaviour of α -nDCG in the two different settings?
- Is the robustness of α -nDCG affected by setting α according to the safe threshold?

Therefore, we shall further investigate the above issues in the next chapter. We analyse the behaviour of α -nDCG under different settings using data from TREC 2009 and 2010 Web Diversity tracks. In particular, submitted runs of TREC systems are re-evaluated and investigated, examining how the variation of α affects the evaluation of document rankings and the subsequent changes obtained in rankings of systems. We also study the intuitiveness of α -nDCG by looking at actual rankings from TREC submissions, compared with user models. Moreover, we generate several simulated

runs to show the consequence of this variation with respect to different levels of ranking performance.

Chapter 8

Evaluation of the Safe Threshold for α -nDCG

8.1 Introduction

α -nDCG is widely used in research and development; however, the effect of its parameter, α , has not yet been thoroughly investigated. We have previously shown that an arbitrary setting of α leads the measure to behave counter-intuitively in particular circumstances. We introduced an approach, providing a solution that determines the parameter of α -nDCG on a query-by-query basis. The approach relies upon imposing the value of α based on the number of sub-topics on each query.

In this chapter, we aim to investigate the effect of setting α according to a common practice, i.e. $\alpha = 0.5$, against that according to our proposed approach. We study the intuitiveness of α -nDCG with different settings by looking at actual rankings from TREC 2009-2010 Web track submissions. By varying α across queries, we further examine whether the reliability of the measure is harmed or not. The discriminative power of α -nDCG is empirically studied using the *paired bootstrap hypothesis test* [Sakai, 2006] together with the *stability* [Buckley and Voorhees, 2000] and the *sensitivity* [Voorhees and Buckley, 2002] measures using the swap method. We can confirm the ability of the measure so as to identify performance differences of distinct retrieval systems, as opposed to differences observed by chance.

Additionally, we aim to study the impact of α 's settings on the base of more comprehensive grounds; however, TREC systems do not represent all scenarios we want

to investigate. In particular, those systems do not achieve very high retrieval performances. To this aim, we employed simulations to generate synthetic system rankings within various performance categories such as high, medium, and low. We analyse and evaluate the simulated systems within each performance category. The experimental results are reported and discussed in this chapter.

This chapter is structured as follows. Section 8.2 outlines experimental plan and research questions investigated in this study. In Section 8.3, we show the results from the studies, followed by the discussion of their analysis in Section 8.4. The chapter concludes in Section 8.5, where we summarise the obtained results and our contributions.

8.2 Experiment and Validation

In the following subsections, we present experimental methodology of the empirical studies. We first define the research questions that our studies want to answer. Then, we define assumptions that will guide the development of the experiments. Finally, we outline the plan of the experiment so as to ensure collected data will address the questions of interest.

8.2.1 Research Questions

The example scenarios of Section 7.5 have been useful to understand the behaviour of α -nDCG with respect to the value of α . A formal threshold we had derived suggests that a value of α should be specified on a query-basis. By re-examining the same scenarios, we had shown how the threshold turns the measure in evaluation to follow the TREC guidelines. Next, we further analyse α -nDCG in realistic scenarios, aiming to answer the following research questions for a qualitative perspective:

- **RQ1:** Can a behaviour similar to that found in the example scenario be exhibited when considering “real” systems, e.g. TREC systems?
- **RQ2:** What happens when the safe threshold approach of Section 7.6 is followed?
- **RQ3:** Do different settings of α lead to different system rankings?

- **RQ4:** Which systems are affected by the difference in the evaluation settings?
- **RQ5:** Are the intuitiveness and the reliability of α -nDCG affected?

8.2.2 Experimental Assumptions

Throughout our study we make a number of working assumptions for the purpose of evaluation to study the intuitiveness and reliability of α -nDCG. In particular,

- 1) We restrict our investigation to *binary judgements*, in line with the TREC 2009, 2010, and 2011 Web Diversity track;
- 2) We consider a *uniform* intent-probability, i.e. the users are equally interested in each one of the identified query-intents;
- 3) We use the user models in Section 7.3 to assume the user in the context of sub-topic retrieval. The user therefore considers systems that retrieve relevant documents covering all query-intents more effective than systems that retrieve only redundant relevant documents without providing a complete coverage of the query; and
- 4) Given the same level of intent-coverage at a specific rank r , we assume that users prefer systems that retrieve more relevant documents to systems that provide the same coverage, but retrieve a lower number of relevant documents.

8.2.3 Plan of Experiments

To answer research questions in Section 8.2.1, we analyse document rankings at rank 10, and mainly focus on two cases: when α is set to 0.5, and when α is set to $st + 0.01$ (and with $\beta = 1$). We use data from the TREC 2009 and 2010 Web Diversity tracks [Clarke et al., 2009a, 2010]. Figure 8.1 reports the sub-topic distribution over the 98 queries¹ contained in the dataset, together with the percentage of queries, i.e. the percentage of topics that exhibit different system rankings when evaluated with the two different settings of α , that are affected by the issue uncovered in Section 7.5. Note

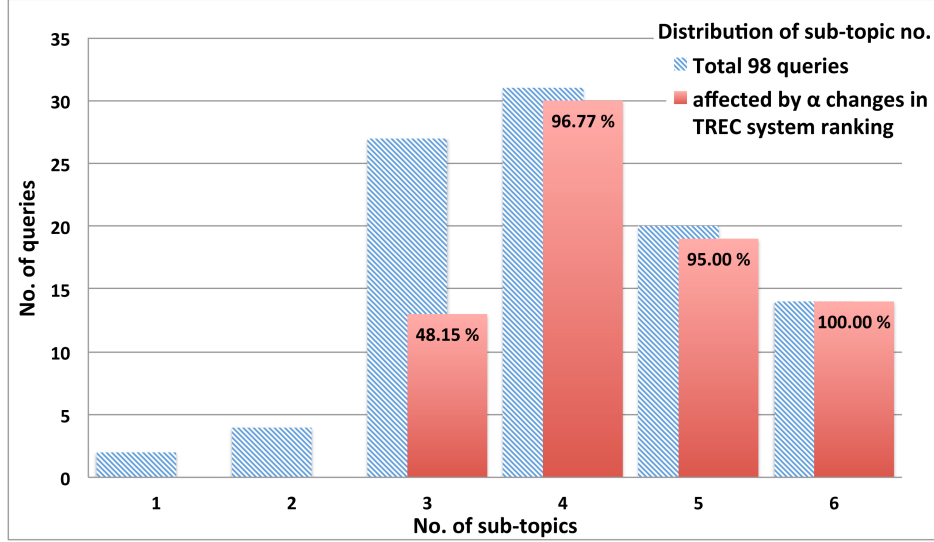


Figure 8.1: The distribution of TREC 2009 and 2010 queries with respect to the number of sub-topics they contain is shown in blue. The relative percentage of queries, for which the setting $\alpha = st + 0.01$ produces different system rankings than the setting $\alpha = 0.5$, is shown in red.

that we excluded from our investigation queries with two or less sub-topics because $\alpha = 0.5 > st$ when $|S| \leq 2$ (i.e. they are not affected by the issue we are investigating).

In Section 8.3.1 we examine a real case example of TREC system runs where common settings of α provide a counter-intuitive system ranking, while $\alpha = st + 0.01$ provides a system ranking consistent with the user model of Section 7.3. The real case example allows us to investigate document lists retrieved by which systems are considered effective for α -nDCG with two different settings. This investigation suggests that α -nDCG with $\alpha = 0.5$ prefers the lists of redundant documents to the lists of documents containing various sub-topics in opposition to $\alpha > st$.

We generalise this finding by examining the kinematics of the system rankings in Section 8.3.2. The kinematics, i.e. how setting $\alpha > st$ modifies the system rankings if they are compared with those obtained with $\alpha = 0.5$, allows us to examine both the amount of differences between system rankings and the positions where the movements take place. This analysis suggests that system rankings formed with $\alpha = 0.5$ are

¹There are two missing queries in TREC 2010, i.e. queries 95 and 100.

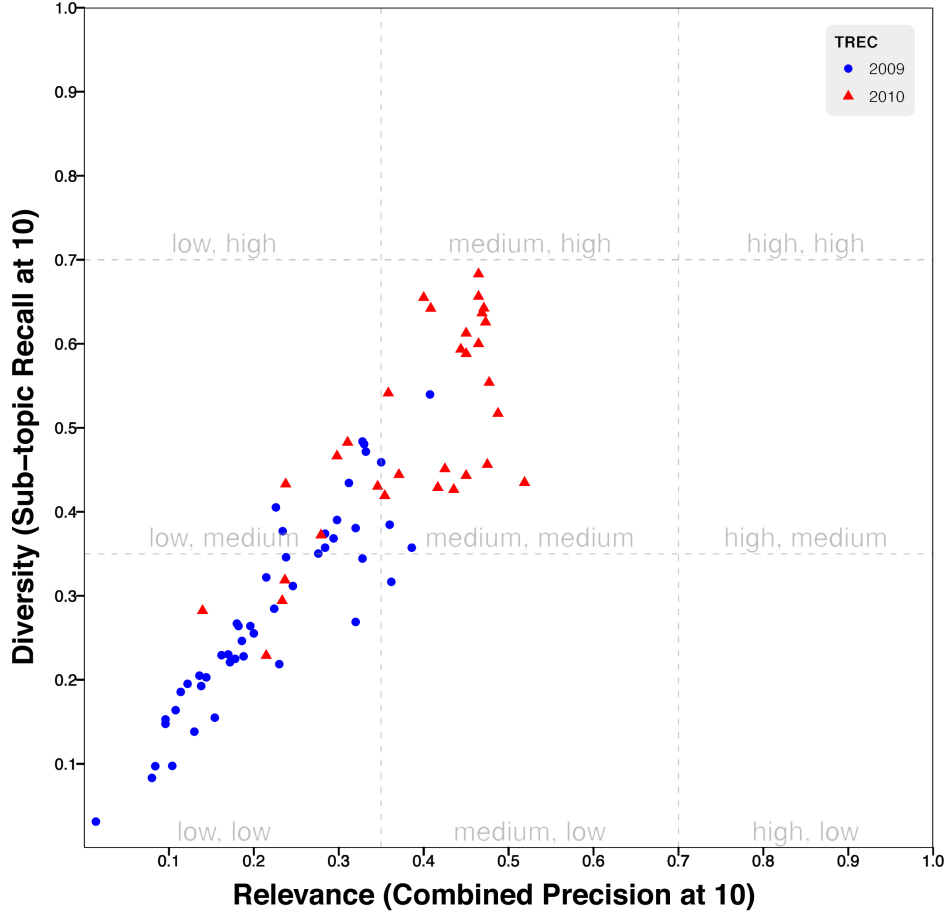


Figure 8.2: Average performance of the systems participating at TREC 2009 and 2010 Web Diversity track, divided into nine performance categories as assessed by the combination of s-recall and combined precision. Note that no system achieves high levels of combined precision and sub-topic recall.

different from those obtained with $\alpha = st + 0.01$: specifically, differences are found in the top positions of the rankings.

The observations suggested by the study of the kinematics are further generalised in Section 8.3.3.1, where we consider *Kendall's* τ rank correlation and *AP correlation* (τ_{ap}) [Yilmaz et al., 2008] between the system rankings obtained with the two different settings of α . In particular, while τ treats discrepancies amongst systems to have equal impact regardless of their positions in the system rankings, τ_{ap} is an *asymmetrical* and

top-heavy coefficient, thus giving higher weight to ranking differences occurring at top ranks. The considerations can be drawn from the initial analysis of 48 and 32 systems submitted to the diversity task of TREC 2009 and 2010 Web tracks, respectively. We categorised these systems into three levels based on combined-precision (*c-precision*) at rank 10 [Chandar and Carterette, 2011; Clarke et al., 2011] and sub-topic recall (*s-recall*) at rank 10. Combined precision at rank r for a query is calculated as the number of documents relevant to *any* sub-topic retrieved up to rank r , divided by r (e.g. $r = 10$). With three levels of *c-precision* and *s-recall*, systems can be divided into nine categories.

Figure 8.2 illustrates *c-precision*@10 vs *s-recall*@10 averaged over all queries for these eighty systems. Systems of TREC 2009 are plotted with blue circles and those of 2010 are red triangles. Note that in TREC 2010 there are two systems, which obtain exactly the same retrieval performances. As shown in the figure, system performances with respect to *c-precision* and *s-recall* are highly correlated. However, none of the systems have contradicting performances, e.g. (high, low) or (low, high), although both situations are theoretically possible. For instance, high *c-precision* and low *s-recall* could be achieved by a system that finds many relevant documents covering few sub-topics, whereas low *c-precision* and high *s-recall* could be achieved by a system that finds a few relevant documents covering many different sub-topics. Moreover, we are most interested in systems that fall into the (high,high) category, but they are under-represented amongst the set of systems that have been used in TREC.

In order to examine the behaviour of α -nDCG on a wider array of systems, we further investigate τ , τ_{ap} and Person’s correlation between system rankings using *synthetic* data, which are generated for each performance-category as identified by *s-recall* and *c-precision* (Section 8.3.3.2).

Finally, in Section 8.3.4 we study the discriminative power of α -nDCG under the two settings of α , so as to assess whether the reliability of the measure is degraded when following our proposal. We confirm our finding regarding the discriminative power by further analysing the stability and sensitivity of α -nDCG.

8.3 Results and Analysis

8.3.1 A Real Case Example

Table 8.1 presents the document rankings and the corresponding relevant sub-topics of the top six systems for query 35 of TREC 2009. The top and bottom rows of the table report the order of the systems according to α -nDCG@10 (whose value is reported in brackets) with $\alpha = 0.5$ and $\alpha = st + 0.01$ (with $st = 0.8$), respectively. Note that when $\alpha = 0.5$, α -nDCG@10 suggests that “uogTrDPCQcdB” is the best performing system as all retrieved documents are relevant. However, these documents cover only two of the six sub-topics. This ranking is highly effective in retrieval tasks such as ad-hoc retrieval, where there is no notion of sub-topics. However, with the TREC guidelines¹ for the Web Diversity task indicating that systems should provide complete sub-topics coverage, while avoiding excessive redundancy. Therefore, the “uogTrDPCQcdB” run is far from being highly effective, as it does not provide a broad coverage of the sub-topics, i.e. it does not diversify the document ranking. On the contrary, the runs identified as “uwgym”, “mudvimp”, and “MSDiv2” should be ranked higher than “uogTrDPCQcdB”, as they cover more sub-topics (i.e. 4, 4, and 5) although retrieving some non-relevant documents. This is because in the diversity retrieval task relevance is not the only evaluation criteria: rankings should also address different sub-topics. While systems such as “uwgym” retrieve less relevant documents than “uogTrDPCQcdB”, they provide a broad coverage of the query’s sub-topics. Whereas “uogTrDPCQcdB” provides a *very relevant* ranking, its results are *not* at all *diverse*.

When our method is used, i.e. α is set as $st + 0.01$, α -nDCG provides a system order in line with the TREC Web Diversity guidelines. In fact, “uwgym”, “mudvimp”, and “MSDiv2” obtain higher scores than “uogTrDPCQcdB”. In particular, “uwgym” is assessed as being the best system for query 35 as it retrieves at rank one a document that covers more sub-topics than that retrieved by “mudvimp” and “MSDiv2”. Similarly “MSDiv2” is ranked lower than “mudvimp”, although it covers 5 sub-topics against the 4 of “mudvimp”. This is because the latter system achieves high sub-topic coverage at earlier ranks, i.e. “MSDiv2” covers 5 sub-topics only at rank 8, while “mudvimp” covers 4 sub-topics after retrieving 5 documents.

¹See Appendix D

Table 8.1: Top six submitted runs from TREC 2009 Web Diversity track on query 35, evaluated by α -nDCG@10 (*Middle*). System rankings with scores when $\alpha = 0.5$ (*Top*) and $\alpha > st$ (*Bottom*).

System ranking according to α -nDCG@10, $\alpha=0.5$						
rank	#1 (0.561)	#2 (0.553)	#3 (0.537)	#4 (0.535)	#5 (0.528)	#6 (0.480)

rank	uogTrDPCQcdB		mudvimp		uwgym		UamsDweblFou		NeuDiv1		MSDiv2	
	Sub-topic		Sub-topic		Sub-topic		Sub-topic		Sub-topic		Sub-topic	
#1	3	6	3	6	3	6	3	6	3	6	2	
#2	3	6	1	4			3	6				
#3	3	6					3	6	3	6	3	6
#4	3	6	3		1		3		3	6		
#5	3	6	2						3	6		
#6	3	6			1				3	6		
#7	3	6	1			5			3	6		
#8	3	6							3	6	1	4
#9	3	6							3	6		
#10	3	6							3	6		

System ranking according to α -nDCG@10, $\alpha = 0.81$ (i.e. $\alpha > st$ and $st = 0.8$ when $\beta = 1$)						
rank	#4 (0.498)	#2 (0.609)	#1 (0.617)	#5 (0.497)	#6 (0.486)	#3 (0.574)

A similar case can be observed when comparing the document orderings of “UamsDwebLFou” and “NeuDiv1”. In particular, “UamsDwebLFou” is ranked higher than “NeuDiv1” even though at rank 10 it retrieves the relevant sub-topics fewer times: seven times against the eighteen times of system “NeuDiv1”. This is because the two sub-topics that are retrieved by “UamsDwebLFou” at rank two provide a higher gain than all the redundant sub-topics retrieved by “NeuDiv1” at later ranks. The gains achieved by “NeuDiv1” amongst ranks 5 to 10 are in fact heavily discounted by both position and redundancy. Note that despite the different settings of α , the relative order of systems “UamsDwebLFou” and “NeuDiv1” in terms of α -nDCG does not change.

8.3.2 Kinematics

The kinematics of the system rankings for TREC 2009 and 2010 are reported in Figures 8.3 and 8.4 respectively. The horizontal axis represents the ranking of systems, and the vertical axis shows the query IDs with the number of sub-topics in brackets. In the figures, a blue cell with cross indicates that the corresponding system is ranked at a lower position when $\alpha > st$ than when $\alpha = 0.5$. Conversely, a red cell indicates that when $\alpha > st$ the corresponding system will go up in the ranking. The analysis of the system rankings’ kinematics gives us insights into the differences between the system orderings generated by the two settings of α . Note that we reports only the queries that there are the disagreements of system rankings. In particular, the kinematics of TREC 2009 systems (and likewise that of TREC 2010 systems) suggests that disagreements (and thus movements) are likely to happen as the number of sub-topics increases, although these also depend on the degree of sub-topic coverage provided by the relevant documents. In fact, more extensive movements are found when considering queries with 5 and 6 sub-topics. Furthermore, it is possible to observe that movements mainly involve top-end systems (i.e. the top 10-20 systems in terms of performance). Whereas, there are only little changes that involve low-ranked systems as these often do not return any relevant documents. For example for query 35, α -nDCG@10=0 for systems ranked between positions 30 and 48. We will further analyse the correlation of two system rankings in various ranking positions, in particular @10, @15, @20, at which the movements are likely to happen.

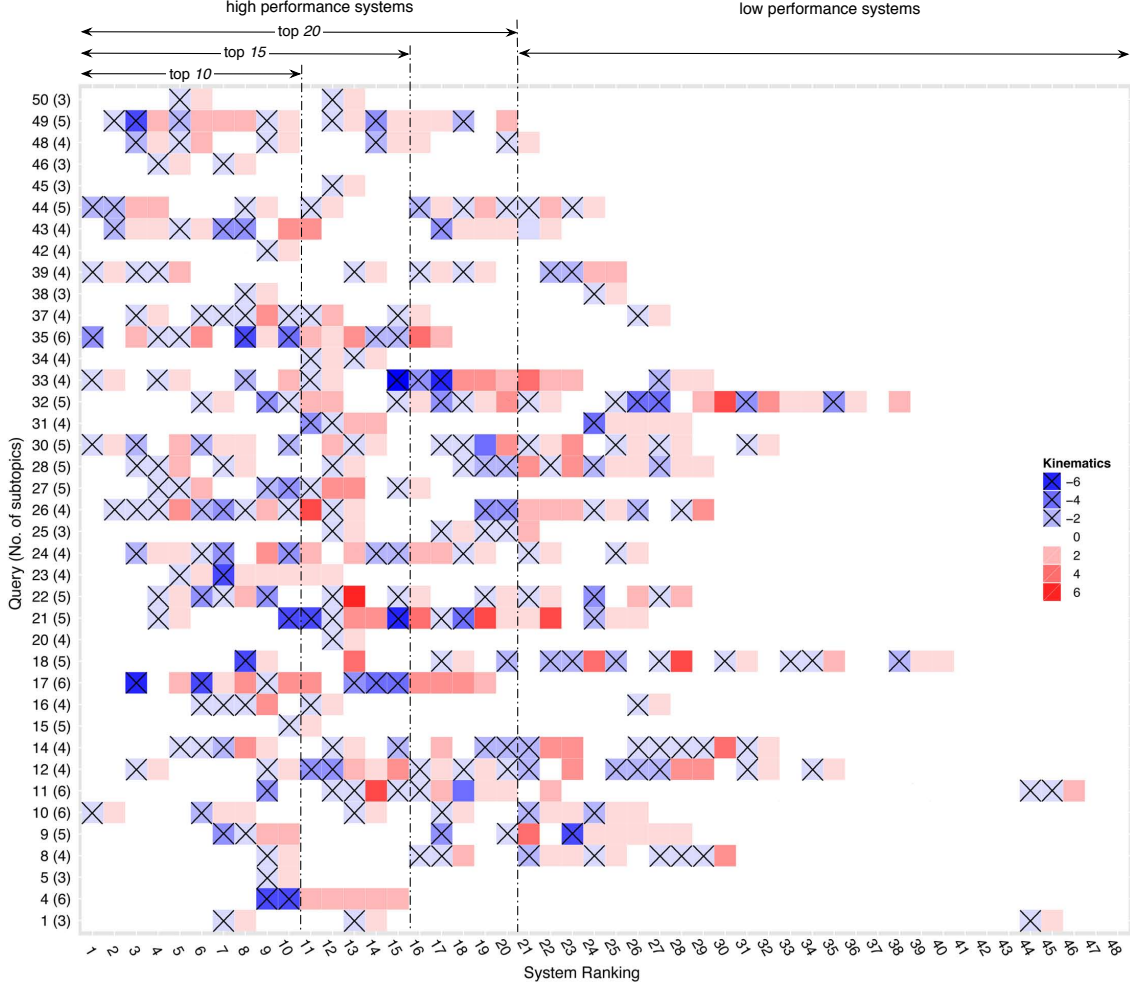


Figure 8.3: Kinematics of 48 system runs submitted to TREC 2009 on 39 queries, with respect to α -nDCG@10 when $\alpha=0.5$, and their movements against α -nDCG@10 with $\alpha > st$.

8.3.3 Correlations

In this section we study the correlations between the system rankings obtained when employing different settings of α -nDCG and when compared with s-recall and s-mrr. Sakai and Song [2011] reported a similar analysis based on correlations between α -nDCG (with $\alpha = 0.5$) and s-recall. These however were computed using system rankings obtained averaging the performances over all the queries contained in the dataset.

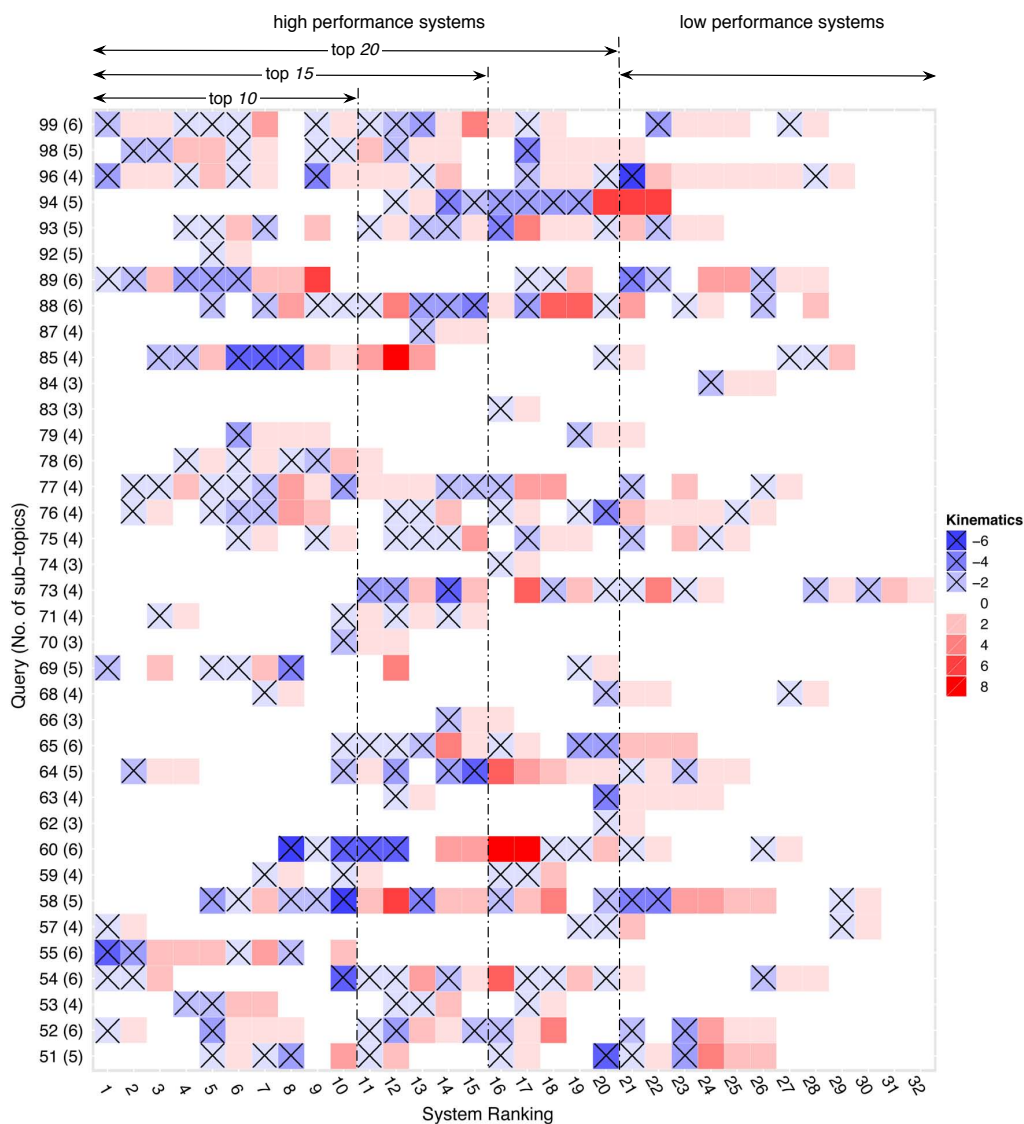


Figure 8.4: Kinematics of 32 system runs submitted to TREC 2010 on 37 queries, with respect to α -nDCG@10 when $\alpha=0.5$, and their movements against α -nDCG@10 with $\alpha > st$.

Here, we instead consider system rankings generated by the different measures on a query-by-query basis, averaging afterwards their correlations over all the queries in the dataset.

Table 8.2: Kendall’s τ and τ_{ap} between rankings of systems submitted to TREC 2009 and 2010 and evaluated with α -nDCG@10 and $\alpha=0.5$ or $\alpha > st$, or s-recall. *All* systems are considered.

TREC 2009	s-recall	$\alpha = 0.5$	$\alpha > st$
$\alpha = 0.5$	0.437/0.371	-	-
$\alpha > st$	0.450/0.395	0.951/0.918	-

TREC 2010	s-recall	$\alpha = 0.5$	$\alpha > st$
$\alpha = 0.5$	0.490/0.406	-	-
$\alpha > st$	0.509/0.426	0.933/0.899	-

8.3.3.1 Real Systems

The analysis of the kinematics reported in the previous section suggested that system rankings obtained employing different settings of α are different. In this section, we provide a measurement of how much they differ, and we compare the rankings with that obtained using s-recall. Table 8.2 reports τ and τ_{ap} between systems ranking obtained with α -nDCG with $\alpha = 0.5$ and $\alpha > st$, and s-recall. The rank correlation analysis reveals that there are *differences* (despite small) between the system rankings, and on average $\tau = 0.951$ and $\tau_{ap} = 0.918$ for TREC 2009, and $\tau = 0.933$ and $\tau_{ap} = 0.899$ for TREC 2010. While these values may suggest that the system rankings are very similar, a further analysis reveals that the systems at the top of the ranking vary considerably when considering $\alpha = 0.5$ or $\alpha > st$. This is evident when examining the results reported in Table 8.3. In fact, when only the top 10 systems are considered, the two system rankings are only weakly correlated (for both τ and τ_{ap} and in both TREC tracks). The correlations increases as the number of considered systems increases: as pointed out for the kinematics, this is often due to poorly performing systems that do not retrieve relevant documents and for which therefore there is no difference in evaluation by the two α -nDCG settings.

We also analysed the correlations between s-recall@10 and the two settings of α -nDCG@10: these are reported in Table 8.2. Both τ and τ_{ap} of rankings obtained using $\alpha > st$ are higher than those obtained with $\alpha = 0.5$, suggesting that our method delivers system rankings that are more adherent to those obtained with s-recall. While this does not guarantee specific advantages, it witnesses that more weight is given to sub-topic coverage, and by reflection to diversity, when considering $\alpha > st$ rather than the common setting.

Table 8.3: Kendall’s τ and τ_{ap} between rankings of systems submitted to TREC 2009 and 2010 and evaluated with α -nDCG@10 and $\alpha=0.5$ or $\alpha > st$, or s-recall. Only the top 10, 15 and 20 systems are considered.

TREC 2009	@10	@15	@20
τ	0.696	0.773	0.831
τ_{ap}	0.742	0.801	0.840

TREC 2010	@10	@15	@20
τ	0.717	0.785	0.818
τ_{ap}	0.686	0.770	0.804

8.3.3.2 Synthetic Systems

To study the impact of α ’s settings on the base of more comprehensive grounds, we investigate a wide array of systems with respect to their performance, as measured by s-recall and c-precision. By doing so, we are able to investigate systems that are under-represented in the real TREC data (see Figure 8.2): in particular we can examine systems that achieve high levels of s-recall and c-precision. To this aim, we employed simulations to generate synthetic system rankings within each performance-category. To obtain the synthetic data, systems were sampled by varying s-recall, c-precision, and document ordering. Our procedure consists of two steps:

Step 1: We used the Fisher-Yates shuffle algorithm [Knuth, 1969, where it is referred to as Algorithm P] to sample documents from the TREC 2009 and 2010 Web Diversity track’s relevance assessments. Subsequently, we generate 10 random samples of system rankings that satisfy each of the experimental conditions: i.e. low, medium, high s-recall@10 and low, medium, high c-precision@10 (these are referred to with labels corresponding to values between (0,0.35] for low, (0.35,0.70] for medium, and (0.70,1] for high).

Step 2: We then re-order the documents appearing in the first 10 sample rankings to ensure that maximum s-recall is obtained at different ranks (i.e. between ranks 1 and 10). From each initial sample we further produce other 10 system rankings obtaining 100 samples for each performance-category. By doing so, we vary the performances of system runs with respect to the rank positions.

Table 8.4: Kendall's τ and τ_{ap} between rankings of systems evaluated with α -nDCG@10 and $\alpha=0.5$ or $\alpha > st$. *Synthetic* systems are considered.

sub-topic recall	<i>all</i> (0.00, 1.00]	0.919/0.914	0.836/0.826	0.800/0.791	
		↑↑	↑↑	↑↑	
	<i>high</i> (0.70, 1.00]	0.923/0.927	0.877/0.879	0.813/0.802	⇒ 0.758/0.759
	<i>medium</i> (0.35, 0.70]	0.941/0.938	0.855/0.857	0.841/0.830	⇒ 0.830/0.842
	<i>low</i> (0.00, 0.35]	0.981/0.985	0.950/0.953	0.936/0.924	⇒ 0.926/0.929
		<i>low</i> (0.00, 0.35]	<i>medium</i> (0.35, 0.70]	<i>high</i> (0.70, 1.00]	<i>all</i> (0.00, 1.00]
		<i>combined precision</i>			

We then evaluate approximately 900 rankings¹ using α -nDCG with $\alpha = 0.5$ and $\alpha > st$. Table 8.4 reports the results obtained by our simulations. Correlations are divided into groups with respect to the performance of the systems they refer to. The correlations of single groups are then reported in the cells of the table. For example, the bottom-leftmost cell of the table refers to systems performing poorly both in terms of s-recall and c-precision. Results are also aggregated by row and by column, representing systems that perform poorly with respect to c-precision, regardless of s-recall (top-leftmost cell of the table).

As for real systems, τ and τ_{ap} give an indication of the differences that are found when evaluating rankings using α -nDCG and $\alpha = 0.5$, and when adopting our method. Results obtained on synthetic data suggest that differences are likely to happen for top ranked systems, i.e. the most effective systems, as correlations are lower for the top-rightmost cells of Table 8.4. Similar conclusions can be derived by examining slices of the table. Consider for example the data obtained when aggregating rows or columns: those referring to the most effective systems present lower correlations

¹Five queries (i.e. 7, 27, 49, 92, and 94) are unable to generate synthetic systems in some performance-categories because relevance judgements do not contain documents for simulating systems in such categories.

Table 8.5: The performance of five synthetic runs on query 21 wrt. α -nDCG with $\alpha = 0.5$ and $\alpha > st$, and maximum s-mrr

synthetic runs	α -nDCG $\alpha = 0.5$	α -nDCG $\alpha > st$	Max(s-mrr)
p70_s70_32	0.427	0.437	0.500
p70_s70_38	0.394	0.380	0.143
p70_s70_39	0.400	0.378	0.125
p70_s70_30	0.401	0.375	0.111
p70_s70_31	0.356	0.353	0.100

than the others. In particular, τ and τ_{ap} in only high c-precision (rightmost cell of the aggregated row) and only high s-recall (top cell of the aggregated column) indicate that the discrepancies of system rankings are more likely to happen amongst systems obtaining high s-recall than systems obtaining high c-precision. This analysis confirms on a larger scale what has been observed in the study of the kinematics and of the correlations of real TREC systems.

Table 8.5 shows five sample simulated runs in category (medium, medium) for query 21 of TREC 2009, and the corresponding scores of α -nDCG with two settings and s-mrr. In the table, we highlighted with a box three systems (i.e. “38”, “39”, and “30”) that are ranked differently depending upon the setting of α -nDCG that is employed. However, while the system ordering between the two settings of α -nDCG differ, setting $\alpha > st$ produces a system ranking that is consistent with that produced if s-mrr is employed. To investigate if this is often the case, we computed the Pearson’s correlation between the scores obtained by s-mrr and those obtained by the two settings of α , in all cases where disagreements between the two settings of α -nDCG occur. The results are reported in Table 8.6. These clearly show that setting $\alpha > st$ promotes systems that do provide complete or broad coverage of the queries’ sub-topics. While, the fact that rankings obtained with s-mrr and $\alpha = 0.5$ are anti-correlated may suggest that this setting of α -nDCG does not reward broad or complete coverage of sub-topics. The α -nDCG with $\alpha = 0.5$ assesses more the retrieval of relevant information, but less the covered sub-topics. This may also indicate that the same situation observed in the real case example, reported in Section 8.3.1, occurs frequently.

Table 8.6: Pearson’s correlation between the system rankings obtained by s-mrr and those obtained by the two settings of α -nDCG.

α -nDCG	Pearson Correlation
$\alpha = 0.5$	−0.427
$\alpha > st$	+0.453

8.3.4 Reliability of α -nDCG

8.3.4.1 Discriminative Power

In the following we attempt to quantify the reliability of α -nDCG under different settings of α . Our goal is to verify that varying α according to our proposal does *not* decrease the reliability of α -nDCG in terms of *discriminative power*. To this aim, we use the method introduced by Sakai [2006] and based on the two-tailed paired bootstrap test. The method involves conducting a statistical significance test for different pairs of experimental runs. In particular, it computes the percentage of pairs that are significantly different at specific fixed significance levels. In our experiments, we use all the systems submitted to the TREC 2009 and 2010 Web Diversity Tracks to generate pairs. Thus we obtained $48 \times (48-1)/2 = 1,128$ pairs for TREC 2009, and $32 \times (32-1)/2 = 496$ pairs for TREC 2010. As a query set, we only considered queries that contained 3 or more sub-topics. As a significance test, we employed a two-tailed paired bootstrap test with 1,000 samples and a fixed significant level of 0.05. The bootstrap samples were obtained by sampling queries *with* replacement.

In our experiment, we found that the two settings of α -nDCG produce slightly different levels of discriminative power. For example, in TREC 2009 the discriminative power of α -nDCG with $\alpha = 0.5$ is 60.72% while that with $\alpha > st$ is 61.08%. The small difference between the discriminative powers obtained by the two settings is not surprising, as our results are based on a comparison of the same measure. However, it can be noticed that setting α according to our method does not harm the discriminative power of α -nDCG. Instead, it slightly improves its reliability.

Since in the previous sections it has been observed that system rankings generated by the different settings of α differ within the top positions, we further analyse the differences in terms of discriminative power by considering only the top 20 systems

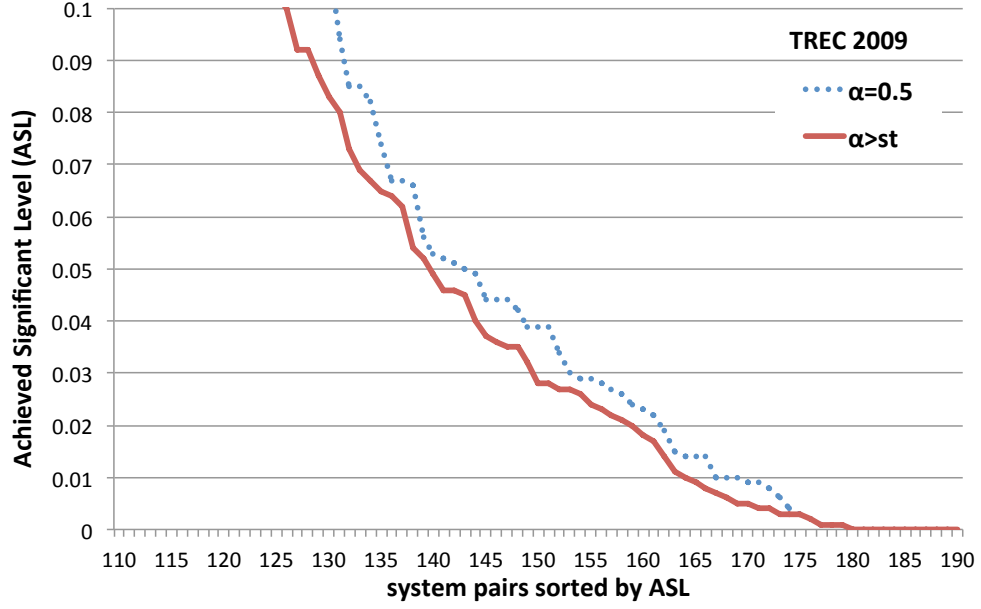


Figure 8.5: ASL curves based on Paired Bootstrap Hypothesis Tests on TREC 2009.

with respect to c-precision and s-recall. This produced $20 \times (20-1)/2 = 190$ pairs to be examined.

Figures 8.5 and 8.6 illustrate the *Achieved Significance Level (ASL)* curves of α -nDCG with the two setting of α for TREC 2009 and 2010. The horizontal axis represents the 190 run pairs sorted in decreasing order of ASL. The vertical axis represents the ASL (i.e. p -value). When considering ASL plots, metrics whose curves are closer to the origin are considered having more discriminative power than the others, i.e. they can detect more significant differences. By examining the plots of Figures 8.5 and 8.6, it can be stated that α -nDCG with $\alpha > st$ is able to discriminate more consistently than α -nDCG with $\alpha = 0.5$. This finding is valid for both TREC 2009 and 2010.

Table 8.7 reports, for two settings of α -nDCG, how many pairs of TREC systems satisfied the condition $ASL < 0.05$. The second column reports the discriminative power while the third reports the estimated difference required for satisfying the condition $ASL < 0.05$. For TREC 2010, the discriminative power of α -nDCG with $\alpha > st$ at 0.05 level is 29.47%. If the difference between the two systems is 0.09 or larger, then the performances of the two systems are significantly different [Sakai, 2006]. The

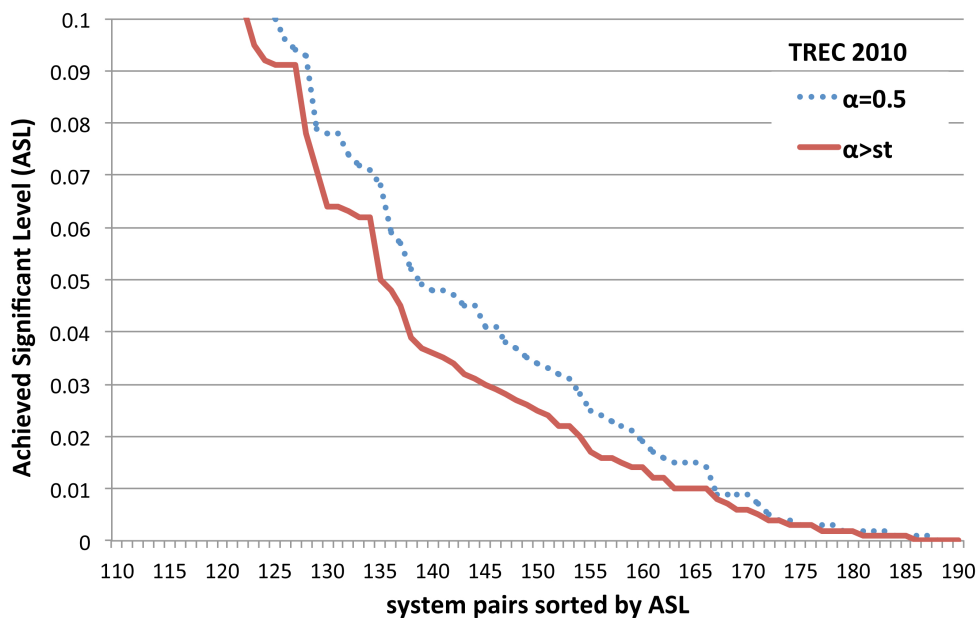


Figure 8.6: ASL curves based on Paired Bootstrap Hypothesis Tests on TREC 2010.

Table 8.7: Discriminative power of traditional metrics at significant level=0.05.

TREC 2009	ASL < 0.05	estimated diff.
$\alpha = 0.5$	47/190=24.73%	0.08
$\alpha > st$	51/190=26.84%	0.08
TREC 2010	ASL < 0.05	estimated diff.
$\alpha = 0.5$	52/190=27.36%	0.09
$\alpha > st$	56/190=29.47%	0.09

comparison between the two setting of α across TREC 2009 and 2010 shows that setting α on a query-by-query basis increases the discriminative power of α -nDCG with respect to the top performing systems.

8.3.4.2 Stability and Sensitivity on Swap Method

To examine the accuracy of the rank correlations between different settings of α -nDCG, we employ the *stability* and *sensitivity* measures based on the *swap* method

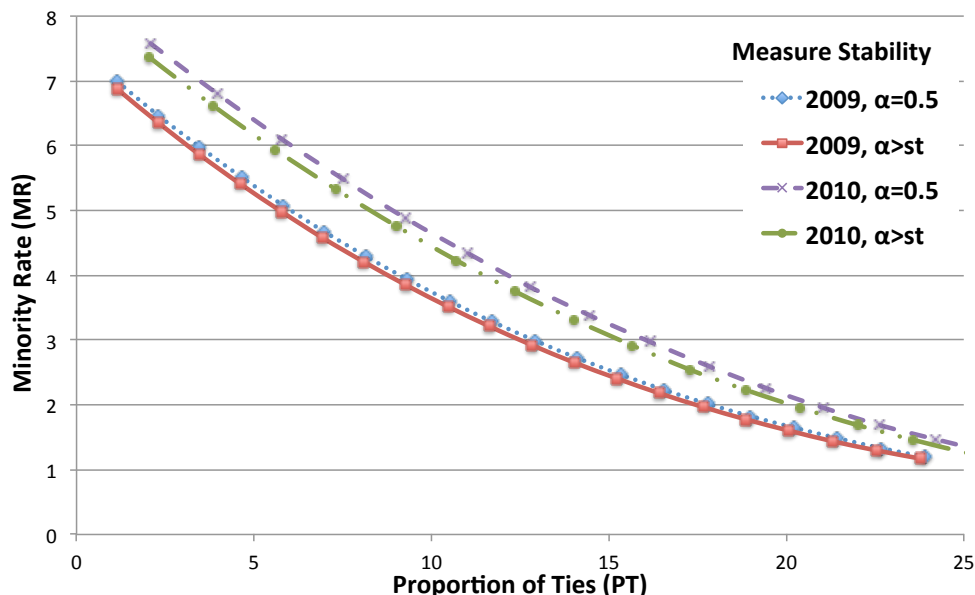


Figure 8.7: MR-PT curves of α -nDCG with the two different settings of α on TREC 2009 and 2010. These curves are used to assess the stability of the measure under the different settings.

proposed by [Buckley and Voorhees \[2000\]](#); [Voorhees and Buckley \[2002\]](#). Unlike Sakai’s bootstrap approach, the swap method is not directly associated with significance tests. It instead relies on a heuristics approach to count the difference in performance between two systems’ pairs. The swap method estimates what the chances are of obtaining a contradictory result from different topic sets (e.g. bootstrap samples): when these are below a specific threshold, a system is considered better than the other. Although the original swap method used sampling *without* replacement, [Sakai \[2006\]](#) suggested that sampling *with* or *without* replacement yields similar results when comparing evaluation measures. In this study, we used the swap method with bootstrap samples¹, and we considered *all* runs submitted to TREC 2009 and 2010 Web Diversity tracks when generating systems pairs.

Figure 8.7 reports the plot of the minority rate (MR) against the proportion of ties (PT) for TREC 2009 and 2010, where the fuzziness values were varied with ($= 0.01, 0.02, \dots, 0.20$) according to [\[Sakai, 2006; Voorhees and Buckley, 2002\]](#). MR repre-

¹We refer the readers to the implementation issues of this techniques to [\[Sakai, 2006\]](#).

Table 8.8: Difference and sensitivity based on the swap method (swap rate $\leq 5\%$) using systems from the TREC 2009 and 2010 Web Diversity tracks.

TREC 2009	Abs. Diff.	Max.	Rel.	Sensitivity
$\alpha = 0.5$	0.15	0.58	25.87%	27.59%
$\alpha > st$	0.15	0.61	24.67%	29.44%

TREC 2010	AbsDiff	Max	Rel	Sensitivity
$\alpha = 0.5$	0.19	0.64	29.63%	17.58%
$\alpha > st$	0.20	0.67	29.81%	18.36%

sents the chance of obtaining a contradictory conclusion given a system pair, whereas PT represents the absence of discriminative power. Thus, a reliable measure is characterised by small values of MR and PT. Furthermore, the closer a curve is to the origin and the better the associated measure is. The results obtained in our analysis are consistent with the finding obtained by the analysis of the discriminative power using the bootstrap test. In particular, α -nDCG with $\alpha > st$ is found to be more stable than the arbitrary setting with $\alpha = 0.5$.

Finally, Table 8.8 summarises the results of the “sensitivity” experiments based on the swap method. Note that the method requires two sets of query samples Q and Q' for estimating the swap rate. We force Q and Q' to be disjoint samples when computing the sensitivity, as in [Voorhees and Buckley, 2002]. The table reports the absolute differences in α -nDCG scores required to have a 5% error rate using topic sets derived from bootstrap samples (“Abs. Diff.”). We also reported the maximum scores recorded amongst all trials (“Max”) and their relative values (“Rel”). The “sensitivity” of a measure is given by the percentage of absolute differences that satisfy the difference-threshold. We used 21 performance-difference bins as suggested by [Voorhees and Buckley, 2002]. The sensitivity results are consistent with those obtained by the bootstrap tests and stability methods. In both TREC 2009 and 2010, setting $\alpha > st$ improves the sensitivity of α -nDCG over the setting $\alpha = 0.5$.

8.4 Findings and Discussion

In this chapter we have investigated a state-of-the-art evaluation measure, α -nDCG, for the TREC Web Diversity task. We evaluated the two different settings of α (i.e. $\alpha = 0.5$ and $\alpha > st$) using real and simulated systems in TREC. The experimental results were analysed based on three different criteria. Firstly, we aimed to investigate the intuitiveness of α -nDCG. To this aim, we observed which document rankings of systems are favoured by α -nDCG in both α settings, and which of these settings are more intuitive with respect to the user models of the diversity retrieval task. Next, we intended to study the effect of the arbitrary setting of α that is employed in common practice. We analysed system rank correlation in order to examine how often and at which rank positions disagreements occur. Finally, we aimed to confirm that setting α according to the safe threshold does not affect the reliability of α -nDCG in terms of its ability to detect performance difference between systems. To this aim, we analysed the discriminative power of α -nDCG as well as its stability and its sensitivity.

For our studies, we derived the following findings that answer research questions

RQ1-5:

- 1) α is not only a user dependent parameter, but it also depends upon the number of query-intents.
- 2) Common settings of α (i.e. $\alpha = 0.5$) prevent α -nDCG from behaving as desired in specific circumstances, i.e. reward systems that provide novel and diversified rankings.
- 3) This issue affects many topics in the TREC 2009 and 2010 Web Diversity tracks and leads to orderings of systems that do not reflect the preferences expressed by user models derived from the TREC Web Diversity task guidelines. In particular, the order of top ranked systems is affected.
- 4) A formal threshold for α can be derived so as to set α on a query-by-query basis guaranteeing that systems are consistently evaluated according to a user model derived from TREC guidelines.
- 5) Empirical evidences suggest that setting $\alpha > st$ improves the reliability and intuitiveness of α -nDCG.

8.5 Summary

In this part, we introduced a theoretically sound approach which derives the safe threshold for α -nDCG on a query-basis. The derivation of our approach suggested that α is not only a user dependent parameter, but also a parameter depending on the sub-topics of a query. We showed the example scenarios that by employing our safe threshold α -nDCG does behave as anticipated by the TREC evaluation guidelines; i.e. “provide complete coverage for a query, while avoiding excessive redundancy”. In this chapter, we found the similar scenarios when considering “real” systems, e.g. TREC systems, as well as “synthetic” systems by simulation. Different settings of α lead to the disagreements of system rankings, in particular in the high performance systems measured by combined-precision and sub-topic recall. The analysis of discriminative power suggests that α -nDCG with $\alpha > st$ is able to detect more significant levels than α -nDCG with $\alpha = 0.5$. This finding was also verified by considering the error rates based on the swap method. The stability and sensitivity of α -nDCG are improved by setting α following the safe threshold.

In summary, by setting α on a query basis according to the safe threshold, the diversity of document rankings can be measured with higher intuitiveness and reliability, without recurring to further modify α -nDCG. Empirical evidences have shown that our method leads to consistently different system rankings when compared to those obtained by setting α according to common practice. Future work will be directed towards examining whether alternative measures, e.g. NRBP [Clarke et al., 2009b], ERR-IA [Chapelle et al., 2009], etc., are affected by similar issues.

Chapter 9

Re-analysing Diversification Approaches for Sub-topic Retrieval

9.1 Introduction

As discussed in the two previous chapters, the common parameter setting of α -nDCG, i.e. $\alpha = 0.5$, as suggested by TREC 2009-2010 Web Diversity track causes the measure to behave counter-intuitively when evaluating IR systems in the context of sub-topic retrieval. Instead of rewarding the systems that provide novel and diversified rankings, α -nDCG over-rates the systems that provide redundant rankings. To avoid such a case, we proposed a query-basis approach that defines the safe-threshold (st) for the parameter α . This problem is crucial since α -nDCG is widely used for training and evaluating experimental systems. The results obtained by the common setting may mislead researchers about the performance of diversification algorithms and then adversely affect the further development of retrieval systems. Similarly, we used α -nDCG with $\alpha = 0.5$ as a measure for evaluation and parameter tuning in the empirical studies conducted in Chapter 6 (Part III). Consequently, the counter-intuitive behaviour of α -nDCG might also affect the results of our studies.

To this aim, this chapter is devoted to re-analysing all the results of ranking strategies considered in the studies. With the safe-threshold proposed, the performances of diversification systems can be truly evaluated and the diverse rankings will be intuitively optimised according to the goal of the diversity task. In the following section, we present the results re-analysed based on our proposal. We then discuss and summarise the obtained results in Section 9.3.

9.2 Results and Analysis

This section shows the effectiveness of diversifying documents as measured in terms of α -nDCG, in particular, when $\alpha > st$. The safe-threshold st is computed for each query according to the equation (7.10). The s-recall (s-r) and s-mrr are also used to evaluate diversified rankings. Table 9.1 reports the re-analysed results of diversification approaches for ImageCLEF 2009 collection, Table 9.2 for TREC ClueWeb 2009 collection, and Table 9.3 for TREC 6,7,8 interactive collection.

Regarding parameters settings, we present the performances of experimental runs that delivered the highest value of α -nDCG@10, averaged over the whole set of query topics in each dataset. The values of parameter settings are shown below the methods. We report the results based on *Ideal Sub-topics*, considered as the upper bound performance that the methods based on sub-topic estimation technique can achieve.

In the tables, the integration approach, $\text{Integr}_{MMR}(\cdot)$, upon different combinations of ranking methods are highlighted by underlining their names. Also, the results obtaining the best performance of the runs regarding given measures are highlighted in bold (excluding three approaches based on the ideal sub-topics). Statistical significant differences (according to a two-tailed t-test, with $p < 0.05$) are analysed against MMR and MPT, and indicated by * and [†] respectively. Note that, in Table 9.3 the analysis of statistical significance is not reported due to the limited number of topics (only 20 topics available) in TREC 6,7,8 interactive sub-topic collection. Thus, computing statistical significance does not convey meaningful information [Bartlett et al., 2001; Voorhees and Harman, 2005].

9.2.1 Re-analysed Results of ImageCLEF 2009

For the ImageCLEF 2009 test collection, the results of our empirical investigation suggest that the instantiations of our integration approach, $\text{Integr}_{MMR}(\cdot)$, outperform those of the inter-dependent document relevance paradigm (i.e. MMR and MPT), with respect to all three measures. With LDA used as a sub-topic estimation technique, the integration approach achieves the best retrieval performance with significant difference against both MMR and MPT in terms of α -nDCG@10. Other sub-topic estimation techniques (i.e. PLSA and K-means clustering) obtain comparable results. These findings suggest that the integration of two retrieval paradigms improves performances in

Table 9.1: Retrieval performances on the *ImageCLEF 2009 (Photo Retrieval)* collection with % of improvement over PRP. Parametric runs are tuned w.r.t. α -nDCG@10 ($\alpha > st$). Statistical significances at 0.05 level against MMR, and MPT are indicated by * and † respectively.

		Models	α -nDCG@10	s-r@10	s-r@20	s-mrr 25%	s-mrr 50%
		PRP	0.4370	0.5330	0.6235	0.7589	0.5221
		MMR ($\lambda = 0.6$)	0.4930 (+12.81%)	0.6761 (+26.86%)	0.7315 (+17.33%)	0.7612 (+0.30%)	0.5341 (+2.31%)
		MPT ($b = 2, \delta^2 = 10^{-2}$)	0.4640 (+6.18%)	0.5688 (+6.72%)	0.6676 (+7.08%)	0.7432 (-2.07%)	0.4996 (-4.31%)
Sub-topic Estimation	K-means	Interp ($\lambda = 1.0$)	0.4370 (0.00%)	0.5330* (0.00%)	0.6235* (0.00%)	0.7589 (0.00%)	0.5221 (0.00%)
		Repre _{PRP}	0.4730 (+8.24%)	0.5701* (+6.97%)	0.6573* (+5.43%)	0.7503 (-1.13%)	0.5173 (-0.92%)
		Integr _{MMR} ($\lambda = 0.7$)	0.5050 [†] (+15.56%)	0.6866 [†] (+28.83%)	0.7501* (+20.31%)	0.7778 (+2.49%)	0.5415 (+3.71%)
	PLSA	Interp ($\lambda = 1.0$)	0.4370 (0.00%)	0.5330* (0.00%)	0.6235* (0.00%)	0.7589 (0.00%)	0.5221 (0.00%)
		Repre _{PRP}	0.4850 (+10.98%)	0.5766* (+8.19%)	0.6805* (+9.15%)	0.7608 (+0.25%)	0.5361 (+2.69%)
		Integr _{MMR} ($\lambda = 0.7$)	0.5160 [†] (+18.08%)	0.6910 [†] (+29.65%)	0.7639 (+22.52%)	0.7883 (+3.88%)	0.5507 (+5.47%)
	LDA	Interp ($\lambda = 1.0$)	0.4370 (0.00%)	0.5330* (0.00%)	0.6235* (0.00%)	0.7589 (0.00%)	0.5221 (0.00%)
		Repre _{PRP}	0.4890 (+11.90%)	0.5683* (+6.62%)	0.6637* (+6.45%)	0.8104* [†] (+6.79%)	0.5406 (+3.55%)
		Integr _{MMR} ($\lambda = 0.9$)	0.5340 * [†] (+22.20%)	0.7100 * [†] (+33.21%)	0.7932 * (+27.22%)	0.8173 * (+7.70%)	0.5589 (+7.05%)
Ideal Sub-topics	Interp ($\lambda = 1.0$)	0.4370 (0.00%)	0.5330* (0.00%)	0.6235* (0.00%)	0.7589 (0.00%)	0.5221 (0.00%)	
	Repre _{PRP}	0.5890* [†] (+34.78%)	0.7901* [†] (+48.24%)	0.8066* [†] (+29.37%)	0.7440 (-1.97%)	0.5544 (+6.18%)	
	Integr _{MMR} ($\lambda = 0.9$)	0.6480* [†] (+48.28%)	0.8136* [†] (+52.65%)	0.8136* [†] (+30.49%)	0.8333* [†] (+9.81%)	0.6301* [†] (+20.69%)	

the case of ImageCLEF 2009. Besides we can notice that the correlation of the diversity performance regarding α -nDCG when $\alpha > st$ to s-recall and s-mrr, is higher than that of α -nDCG when $\alpha = 0.5$ (see Table 6.3). That is, all evaluation measures are in agreement when rating the system performance.

By comparing two ranking paradigms, we see that only the MMR of inter-dependent document relevance paradigm outperforms the two approaches of sub-topic aware paradigm, i.e. the interpolation approach, Interp(.) and the cluster representative approach, Repre_{PRP}(.). Meanwhile, although MPT fails to improve diversification performance over Repre_{PRP}(.), it still performs better than the PRP baseline and Interp(.). It is interesting to notice that the best value of α -nDCG@10 for Interp(.) is obtained

when $\lambda = 1.0$, in which the obtained ranking is equivalent to that of PRP. This suggests that $\text{Interp}(\cdot)$ ranking formula reduces to the PRP one, when parameters are tuned in order to optimise α -nDCG@10 on the whole topic set for ImageCLEF 2009 collection.

Recall that the integration approach, $\text{Integr}_{MMR}(\cdot)$, is inherit from the cluster representative approach, $\text{Repre}_{PRP}(\cdot)$, where $\text{Integr}_{MMR}(\cdot)$ employs MMR to select documents within clusters instead of using PRP. The re-analysed results maintain to suggest that the performances of $\text{Integr}_{MMR}(\cdot)$ runs consistently outperforms those of $\text{Repre}_{PRP}(\cdot)$ in all investigated measures. This phenomenon supports that the integration with MMR for document selection increases diversification performance from the use of PRP.

9.2.2 Re-analysed Results of TREC ClueWeb 2009

In the Table 9.2, we report the result diversification performance on TREC ClueWeb 2009 dataset. As we can see, the run of MPT performs the best in terms of α -nDCG@10, s-recall@10 and @20, but no significant difference has been observed against MMR. When considering s-mrr, all runs based on $\text{Interp}(\cdot)$ suggest that they can cover 25% of the total number of sub-topics earlier than other approaches and are ranked the third for 50% of sub-topic coverage. However, no matter what the parameter is given, the maximum performances of $\text{Interp}(\cdot)$ are limited to 0.1290 regarding α -nDCG@10.

Although the integration approach, $\text{Integr}_{MMR}(\cdot)$, does not perform the best on TREC ClueWeb 2009 dataset, this is due to the fact that the sub-topic estimation technique fails to model sub-topics corresponding to user information needs (compared with the results of $\text{Repre}_{PRP}(\cdot)$). However, the $\text{Integr}_{MMR}(\cdot)$ runs show the consistent improvement over the $\text{Repre}_{PRP}(\cdot)$ in all three sub-topic estimation techniques, i.e. K-means, PLSA, and LDA. Furthermore, If the ideal sub-topic estimation is considered, the $\text{Integr}_{MMR}(\cdot)$ is confirmed to shows the potential to improve the results of the $\text{Repre}_{PRP}(\cdot)$ and outperforms those of the other state-of-the-art approaches, e.g. MMR and MPT. These findings again suggests that with sub-topic evidences MMR can enhance the performance of diversification over PRP by including document dependencies for document selection.

9.2 Results and Analysis

Table 9.2: Retrieval performances on the *TREC ClueWeb 2009* collection with % of improvement over PRP. Parametric runs are tuned w.r.t. α -nDCG@10 ($\alpha > st$). Statistical significances at 0.05 level against MMR, and MPT are indicated by * and † respectively.

		Models	α -nDCG@10	s-r@10	s-r@20	s-mrr 25%	s-mrr 50%
		PRP	0.0590	0.1606	0.2719	0.1787	0.0953
		MMR ($\lambda = 0.6$)	0.1130 (+91.53%)	0.1669 (+3.92%)	0.2771 (+1.91%)	0.1801 (+0.78%)	0.0991 (+3.99%)
		MPT ($b = -5, \delta^2 = 10^{-4}$)	0.1690 (+186.44%)	0.2676* (+66.64%)	0.3486* (+28.20%)	0.2179 (+21.90%)	0.1264 (+32.69%)
Sub-topic Estimation	K-means	Interp ($\lambda = 0.3$)	0.1290 (+118.64%)	0.1721 [†] (+7.16%)	0.2390 [†] (-12.10%)	0.3247* (+98.49%)	0.1410 (+47.95%)
		Repre _{PRP}	0.1350 [†] (+128.81%)	0.1819 [†] (+13.29%)	0.2466 [†] (-9.32%)	0.2077 (+16.21%)	0.1145 (+20.21%)
		Integr _{MMR} ($\lambda = 0.8$)	0.1480 (+150.85%)	0.2038 (+26.90%)	0.2850 [†] (+4.82%)	0.3120 (+74.59%)	0.1366 (+43.34%)
	PLSA	Interp ($\lambda = 0.4$)	0.1290 (+118.64%)	0.1721 [†] (+7.16%)	0.2390 [†] (-12.10%)	0.3247* (+98.49%)	0.1410 (+47.95%)
		Repre _{PRP}	0.1410 (+138.98%)	0.1876 (+16.81%)	0.2858 (+5.10%)	0.2265 (+26.73%)	0.1120 (+17.55%)
		Integr _{MMR} ($\lambda = 0.8$)	0.1582* (+168.14%)	0.2130 (+32.63%)	0.3078 (+13.20%)	0.3230* (+80.75%)	0.1574* (+65.16%)
	LDA	Interp ($\lambda = 0.3$)	0.1290 (+118.64%)	0.1721 [†] (+7.16%)	0.2390 [†] (-12.10%)	0.3247* (+98.49%)	0.1410 (+47.95%)
		Repre _{PRP}	0.1520 (+157.63%)	0.2047 (+27.46%)	0.2902 (+6.74%)	0.2134 (+19.40%)	0.0990 (+3.93%)
		Integr _{MMR} ($\lambda = 0.8$)	0.1684 (+185.42%)	0.2370 (+47.57%)	0.3167 (+16.48%)	0.3281 (+83.60%)	0.1623* (+70.30%)
Ideal Sub-topics	Interp ($\lambda = 0.3$)	0.1290 (+118.64%)	0.1721 [†] (+7.16%)	0.2390 [†] (-12.10%)	0.3247* (+98.49%)	0.1410 (+47.95%)	
	Repre _{PRP}	0.2210* (+274.58%)	0.3332* (+107.53%)	0.3872* (+42.42%)	0.2868* (+60.48%)	0.1780* (+86.85%)	
	Integr _{MMR} ($\lambda = 0.2$)	0.2513* (+325.93%)	0.3517* (+118.99%)	0.3917* (+44.06%)	0.4127* [†] (+130.95%)	0.1915* (+100.94%)	

For the agreement amongst different evaluation measures, the results in TREC ClueWeb collection slightly differ from those in ImageCLEF 2009 collection. The diversity performance regarding α -nDCG correlates to s-recall, but not highly correlate to s-mrr. Note that, however, different measures assess different aspects of rankings. For example, consider the run that performs well with respect to α -nDCG@10 but badly with respect to s-mrr 25%. It can be interpreted that a given system is able to provide a diverse ranking, which contains a large number of relevant sub-topics *within* the top ten documents, but cover 25% of the total number of sub-topics at lower positions.

Table 9.3: Retrieval performances on the *TREC 6,7,8 interactive* collection with % of improvement over PRP. Parametric runs are tuned w.r.t. α -nDCG@10 ($\alpha > st$).

		Models	α -nDCG@10	s-r@10	s-r@20	s-mrr 25%	s-mrr 50%
		PRP	0.4120	0.3868	0.5319	0.2877	0.1618
		MMR ($\lambda = 1.0$)	0.4120 (0.00%)	0.3868 (0.00%)	0.5319 (0.00%)	0.2877 (0.00%)	0.1618 (0.00%)
		MPT ($b = -3, \delta^2 = 10^{-2}$)	0.4410 (+7.04%)	0.4195 (+845%)	0.5523 (+3.84%)	0.3119 (+8.41%)	0.1693 (+4.64%)
Sub-topic Estimation	K-means	Interp ($\lambda = 1.0$)	0.4120 (0.00%)	0.3868 (0.00%)	0.5319 (0.00%)	0.2877 (0.00%)	0.1618 (0.00%)
		Repre_{PRP}	0.2790 (-32.28%)	0.2517 (-34.94%)	0.3483 (-34.52%)	0.1340 (-53.43%)	0.0692 (-57.24%)
		Integr_{MMR} ($\lambda = 0.8$)	0.2971 (-27.89%)	0.2812 (-27.30%)	0.3516 (-33.90%)	0.1547 (-46.23%)	0.0713 (-55.93%)
	PLSA	Interp ($\lambda = 1.0$)	0.4120 (0.00%)	0.3868 (0.00%)	0.5319 (0.00%)	0.2877 (0.00%)	0.1618 (0.00%)
		Repre_{PRP}	0.3073 (-25.41%)	0.3132 (-19.03%)	0.4090 (-23.11%)	0.1788 (-37.84%)	0.0688 (-57.47%)
		Integr_{MMR} ($\lambda = 0.6$)	0.3137 (-23.86%)	0.3178 (-17.84%)	0.3953 (-25.68%)	0.1797 (-37.54%)	0.0657 (-59.40%)
	LDA	Interp ($\lambda = 1.0$)	0.4120 (0.00%)	0.3868 (0.00%)	0.5319 (0.00%)	0.2877 (0.00%)	0.1618 (0.00%)
		Repre_{PRP}	0.3314 (-19.56%)	0.3078 (-20.44%)	0.4049 (-23.87%)	0.2043 (-28.99%)	0.1024 (-36.69%)
		Integr_{MMR} ($\lambda = 0.3$)	0.3541 (-14.05%)	0.3386 (-12.46%)	0.4193 (-21.17%)	0.2253 (-21.69%)	0.1161 (-28.24%)
Ideal Sub-topics		Interp ($\lambda = 1.0$)	0.4120 (0.00%)	0.3868 (0.00%)	0.5319 (0.00%)	0.2877 (0.00%)	0.1618 (0.00%)
		Repre_{PRP}	0.5190 (+25.97%)	0.5664 (+46.41%)	0.6761 (+27.12%)	0.2898 (+0.74%)	0.1575 (-2.67%)
		Integr_{MMR} ($\lambda = 0.9$)	0.5315 (+29.00%)	0.5692 (+47.15%)	0.6793 (+27.72%)	0.2971 (+3.28%)	0.1565 (-3.28%)

9.2.3 Re-analysed Results of TREC 6,7,8 Interactive

Now let us look at the results in TREC 6,7,8 interactive dataset. The results of MPT outperform all the other experimental approaches in all three measures. Furthermore, techniques for sub-topic modelling appear to provide the weak sub-topic evidences to support the approaches of sub- topic aware paradigm and also affect our integration approach. As a result, the Integr_{MMR}(.), Repre_{PRP}(.), or Integr_{MMR}(.) performs as good as or worse than MMR, MPT, and even the PRP baseline. We can see that the results of MMR and Interp(.) are obtained when their hyper-parameter $\lambda = 1.0$, that is, when their ranking formula is equivalent to the one of the PRP. This means that the diversification components in their functions do not provide any useful evidence for promoting diversity. Nevertheless, when sub-topics are estimated from the relevance judgements,

as in the case of the ideal sub-topics, the instantiations of $\text{Repre}_{PRP}(\cdot)$ and $\text{Integr}_{MMR}(\cdot)$ outperform any other approach.

Similar results are found when comparing $\text{Repre}_{PRP}(\cdot)$, or $\text{Integr}_{MMR}(\cdot)$. In all cases, using MMR instead of PRP for document selection improves the effectiveness of document ranking diversification. Moreover, we found a similar trend in the correlation between α -nDCG and the other two measures, i.e, s-recall and s-mrr. The results in the TREC 6,7,8 interactive collection report that all three measures are somewhat in agreement when evaluating the systems.

9.3 Discussion and Conclusion

In this chapter, the results of the empirical investigation were re-evaluated, in particular, using α -nDCG with $\alpha > st$. The analysis we obtained showed similar outcomes to those reported in Chapter 6, but with clearer evidence when comparing them amongst different evaluation measures. That is, all three measures are correlated to each other or show agreement on evaluating system rankings. Furthermore, from tunable approaches, e.g. MMR, MPT, we obtained different diverse rankings optimised with respect to α -nDCG when $\alpha > st$. In the following, we summarise our findings that answer the research questions of experiments previously given in Section 6.2.1.

- 1) Regardless of the runs generated by ideal sub-topics, the inter-dependent document relevance paradigm tends to provide better diversification performance than the sub-topic aware paradigm. In most cases of the experimental runs on *three* test collections, the performance of diversification with MMR and MPT is greater than that with the $\text{Interp}(\cdot)$ and $\text{Repre}_{PRP}(\cdot)$. Within three investigated sub-topic estimation techniques, LDA appears to provide the best support of sub-topic evidences for document diversification.
- 2) As we can see, the integration of two ranking paradigms improves the effectiveness of document diversification. However, the obstacle to achieve the highest performance of the integration approach lies in the technique of modelling sub-topics. This is obviously shown in the results of TREC ClueWeb 2009 and TREC 6,7,8 interactive datasets. All K-means clustering, PLSA, and LDA do

not provide good sub-topic evidences for document diversification (see the poor performance of *Interp(.)* and *Repre_{PRP}(.)* runs).

- 3) When considering both the integration approach and the approaches of sub-topic aware paradigm, there has been a noticeable trend towards increasing performance in terms of α -nDCG@10. The maximum gains against the PRP baseline that the integration approach can potentially achieve are 48.28%, 325.93%, and 29.00% in ImageCLEF 2009, TREC ClueWeb 2009 and TREC 6,7,8 interactive datasets, respectively.
- 4) LDA performs the best in three sub-topic estimation techniques, followed by PLSA and K-means clustering. In our diversification framework, applying MMR for document selection increases an average¹ of 8.09% , 11.58%, and 4.46% over the cluster representative approach of sub-topic aware paradigm, *Repre_{PRP}(.)*, in three test collections, i.e. ImageCLEF 2009, TREC ClueWeb 2009, and TREC 6,7,8 interactive, respectively.
- 5) From the analysis of the results of ideal sub-topics, the integration approach has the potential to increase the performance of sub-topic aware paradigm. It is, however, noted that both the integration approach and the sub-topic aware paradigm rely on the outputs of sub-topic modelling techniques, which have implications on the effectiveness of document diversification using *Interp(.)*, *Repre_{PRP}(.)*, or *Integr_{MMR}(.)*. Consequently, we need to find a robust technique that can effectively model sub-topics, corresponding to user information needs.

To sum up, overall approaches derived from the inter-dependent document relevance paradigm, MMR and MPT, outperform approaches derived from the sub-topic aware paradigm. This is opposite to the results given when analysing by α -nDCG with $\alpha = 0.5$. Amongst the techniques for estimating sub-topics, LDA has been shown to model sub-topics most effectively. However, all the techniques for sub-topic estimation fail to some degree to provide good sub-topic evidences in the case of the TREC ClueWeb 2009 and the TRCE 6,7,8 interactive collections. Unlike the captions of images usually annotated by keywords, the documents derived from web pages and

¹It is an average over three sub-topic modelling techniques, K-means clustering, PLSA, and LDA.

newswire articles naturally contain a lot of noise or spam, affecting the quality of sub-topic modelling. Notwithstanding the integration approach, which combines two ranking paradigms for ranking diversification, has been shown to outperform state-of-the-art approaches, in particular when sub-topics are constructed from the relevance judgements. Therefore, the integration approach has the potential to improve sub-topic retrieval performances when effective topic estimation is deployed. Further investigation will be directed towards the empirical validation of effective topic estimation techniques.

Part V

Conclusion

Chapter 10

Conclusions

The broad objective of this thesis was an exploration into three aspects of diversity-based document retrieval. In Part II, we examined the need for result diversity from the users' perspective. We verified this by a user-centred evaluation of the diverse recommendations mined from users' implicit relevance feedback. In Part III, we introduced a diversification framework for integrating two ranking paradigms for diversity-based retrieval and conducted an empirical experiment on three standard test collections for comparing systems' performance. In Part IV, we introduced a query-based approach to parameterise the de-facto standard measure, α -nDCG, for evaluating diversity and redundancy in search result. To examine the measure's validity, we assessed document rankings obtained from both real and synthetic systems using α -nDCG with our proposed setting as well as other measures. Further, we analysed the reliability of α -nDCG in terms of discriminative power, stability, and sensitivity.

First, this chapter summarises the findings and success of this thesis. Next, the contributions of this thesis are listed in Section 10.2. Finally, we discuss avenues of future research that could complement the works described within this dissertation.

10.1 Summary of Work and Discussion

10.1.1 Diversity-Based Recommender System

Part II of this thesis concentrated on studying the benefits of diversity in search result from real users' perspective. While the need for IR systems that diversify search results have been discussed by the research community, no previous study has supported this

need considering real user interactions and preferences. The basic hypothesis was that users would prefer systems that provide diverse results when they have a multi-aspect information need. To evaluate this hypothesis, we conducted a user study employing simulated work task situations, where users had to carry out searches for gathering multiple aspects of a search topic. In such situations, users would consider and explore information in many aspects before settling on a final selection that satisfies their needs.

As introduced in Chapter 3, the Ostensive Browser Plus (OBP) system was employed for the purpose of this study. OBP is a content-based image browsing system, which visualises user interactions into a graph of user browsing trials and adaptively tailors search results to the user's evolving information need. In addition to content-based browsing, OBP is featured with diverse recommendations and aspectual browsing interfaces. The recommendation functionality retrieves documents that are relevant and diverse in terms of various aspects of image contents. Implicit relevance feedback extracted from user browsing trails are exploited to generate a set of potentially relevant images to recommend. A clustering algorithm is applied on different visual features in order to select and diversify recommended documents. The aspectual browsing interface is implemented on the basis of self-organising exploratory search systems. The interface is composed of multiple independent browsing spaces, by which users can organize their searching process and the consequent results. In general, the OBP system is designed to support complex and exploratory search needs that can be defined and structured by users.

In Chapter 4 we examined the advantages of diversity from the point of view of users. We highlighted that result diversity is useful to support users in exploratory search tasks. Users discovered and defined more aspects when using the OBP system that features diverse recommendations. This evidence is also supported by the analysis of questionnaires answered by participating users. Based on the analysis of user's feedback and satisfaction with the system, it revealed that diverse recommendations are effective in supporting users to find relevant images covering multiple aspects. Furthermore, the recommendation feature helps users find more relevant images for the tasks. Thus, we conclude that implicit relevance feedback can be effectively employed in the image retrieval domain to recommend images relevant to a user's information needs.

Additionally, the results obtained in this part of the thesis suggest that diversity brings substantial potential benefits to users, in particular when they have a multi-aspect information need. Therefore, IR systems developed to serve such needs should consider *relevance* and *diversity* of information, which together reflect the usefulness of the systems as perceived by users.

10.1.2 An Integration Framework for Result Diversification

In Part III we focused on automatic methods for document ranking in diversity retrieval and studied the dominant ranking strategy, the probability ranking principle (PRP). We argued that in the evaluation context of diversity retrieval, the independence assumption of PRP is not upheld. This is because in diversity retrieval, documents' relevance is considered dependent on that of other documents. We analysed several ranking approaches, alternative to PRP. These approaches relax some of the assumptions of PRP by considering dependencies between documents in a ranking. We argued that these alternative approaches can actually be divided into two categories according to their distinct ranking patterns:

- 1) inter-dependent document relevance paradigm;
- 2) sub-topic aware paradigm.

These two patterns can be thought of as two faces of the same coin, as they both aim to promote diversity of relevant sub-topics in a document ranking but they do so in two different ways. In the inter-dependent document relevance paradigm, result diversity is achieved by considering dependencies between documents and promoting documents that differ from each other. Specifically, we examined two general parametric strategies, i.e. maximal marginal relevance (MMR) and modern portfolio theory (MPT) for IR. In contrast, the sub-topic aware paradigm directly models sub-topics from documents, modelled by clusters of similar documents. Thus, sub-topic diversification can be achieved by retrieving documents belonging to different clusters. In particular, we examined three topic modelling techniques for clustering documents such as K-means clustering, latent Dirichlet allocation (LDA), and probabilistic latent semantic analysis (PLSA). Further we analysed two ranking strategies for post-clustering such as the interpolation approach (Interp(.)) and the cluster representative approach (Repre_{PRP}(.)).

In Chapter 5.1 we proposed a general diversification framework, which enables the development of a variety of algorithms for integrating the two ranking paradigms. We showed how our framework can be instantiated for ranking documents. In general, ranking documents following our integration framework can address both inter-dependent document relevance and sub-topic diversity. We posit that those two paradigms can be employed together to improve the effectiveness of ranking diversification, as measured in terms of both relevance and diversity. This is because sub-topics are explicitly estimated and documents are dependently ranked.

In Chapter 6, we conducted thorough empirical experiments on three experimental datasets to observe the performance of various diversity-based ranking approaches. The results of experiments were analysed using standard evaluation measures for diversity retrieval, i.e. s-recall, s-mrr, α -nDCG with a common parameter setting ($\alpha = 0.5$). The findings of our empirical investigation showed that the inter-dependent document relevance paradigm (i.e. MMR, MPT) tends to perform empirically better than the sub-topic aware paradigm (e.g. Interp(LDA) and Repre_{PRP}(LDA)) for diversifying documents. When comparing between different topic modelling techniques, LDA appears to provide the best support of sub-topic evidences for document diversification, where these evidences are the groups of similar documents describing the same sub-topic. When analysing the results of our integration framework, we found a noticeable trend towards better performance increasing from the approaches based on sub-topic aware paradigm. However, the improvements that we witnessed in this evaluation context suggest that our integration approach relies on the techniques used for modelling sub-topics, i.e. K-means clustering, LDA, and PLSA. That is, if those techniques could estimate sub-topics corresponding to multi-intent information needs, our integration approach would have effectively performed ranking diversification and would potentially reach the upper-bound performance of runs generated by the ground truth judgements of sub-topic relevance.

10.1.3 Query-Basis Approach to Derive Safe Threshold of α -nDCG

In Part IV, we examined the intuitiveness of evaluation measures in the context of diversity-based retrieval. We identified an issue with novelty-biased discounted cumulative gain (α -nDCG); the common setting of its parameter α causes the measure to

not behave as desired in specific circumstances. We argued that when α is set to 0.5, α -nDCG excessively rewards systems that *redundantly* cover only a few sub-topics. We showed that this issue is very crucial as it highly influences the measurement of the effectiveness of top ranked systems. In particular, when using the measure as an objective function for learning-to-rank, α -nDCG with a common setting (i.e. $\alpha = 0.5$) will result in producing a document ranking that does *not* correspond to ranking preferences as expressed by possible user models for the diversity task.

In Chapter 7, we proposed a theoretically sound solution by defining a safe threshold for α on a per query basis. The key of our approach is to resolve the parameter setting of α -nDCG with respect to the number of sub-topics present in each query. Our derivation of the safe threshold exposes the fact that α is not only a user-oriented parameter, but also the parameter dependent on the number of sub-topics. Although α -nDCG is devised on the basis of *redundancy*, we analytically discussed that using α -nDCG set according to our safe threshold allows to evaluate the *diversity* of sub-topics in a document ranking.

Afterwards, we examined the impact on evaluation of arbitrary setting α to 0.5. We analysed the behaviour of α -nDCG when evaluating actual document rankings from TREC 2009 and 2010 Web track submissions. We observed how the variation of α affects the evaluation of *document rankings* and the subsequent changes obtained in *the ranking of systems*. We also studied the intuitiveness of the measure by comparing actual document rankings, which are rated high by α -nDCG in two parameter settings, with user preferences defined by the user models as shown in Section 7.3. Following our view of user models for diversity retrieval, it is assumed that users would prefer to first cover all relevant sub-topics; then they prefer to examine *redundant relevant* documents (despite unfavourable) rather than *non-relevant* documents.

Furthermore, we generated synthetic system rankings within different performance-categories so as to thoroughly investigate the impact of α 's setting. We showed that different settings of α lead to disagreements in system rankings, in particular when examining the best performing systems as measured by combined-precision and sub-topic recall. Moreover, by varying α across queries, we examined whether the reliability of the measure is harmed or not. By doing this, we analysed the discriminative power, stability, and sensitivity of α -nDCG in two different settings. Results suggest

that setting α according to our safe threshold does not harm but instead increase the reliability of α -nDCG.

Finally, we re-analysed the results of our experiments in Chapter 6. We observed higher correlation between the results evaluated by three different measures, i.e. s-recall, s-mrr, and α -nDCG with $\alpha > st$. They all mostly agree on the system ratings, in particular the ones that were ranked high for result diversification. The re-analysed results were reported in Chapter 9.

10.2 Contributions

Several contributions emerge within this thesis:

- **A diversity-based recommender system for studying the benefits of result diversification.** We introduced a recommendation approach that mines users implicit relevance feedback to generate relevant results diversified based on image content. The aspectual browsing interface is firstly introduced in the content-based browsing system. The system provides various facilities that can be employed by users to explore data collections, discover various search aspects, and organise their searching process and results (Chapter 3).
- **An understanding of the need for the development of result diversification in IR.** A user experiment provides new insights into the importance of diversity in search result from the point of view of the users (Chapter 4).
- **A framework for integrating two ranking paradigms in result diversification.** We explored and analysed a number of diversification approaches in IR, and categorised them based on their ranking patterns. Our diversification framework enables the development of new ranking strategies, addressing together inter-dependent document relevance and sub-topic diversity (Chapter 5).
- **Empirical experiments of diversification approaches on three standard datasets.** To compare and contrast our diversification framework and other ranking approach, we conducted system-oriented experiments on three test collections: ImageCLEF 2009, TREC ClueWeb 2009, and TREC 6, 7, 8 interactive datasets.

These experiments led to improvements in retrieval effectiveness in terms of relevance and diversity (Chapter 6 and 9).

- **Mathematical and behavioural analysis of α -nDCG.** We showed that an arbitrary setting of α (i.e. $\alpha = 0.5$) leads the measure to behave counter-intuitively, i.e. excessively reward the systems that repeatedly return redundant sub-topics. Our analysis provides a thorough investigation of the measure’s behaviours in evaluating diversified rankings (Chapter 7).
- **A theoretically sound approach to determine a safe threshold for α on a per-query basis.** We prove that by employing our safe threshold, α -nDCG is more adherent to the TREC guidelines; i.e. “provide complete coverage for a query, while avoiding excessive redundancy” (Chapter 7).
- **Thorough experiments of the intuitiveness and reliability of α -nDCG.** To demonstrate the validity of the proposed measure, we conducted experiments on both real and synthetic document rankings and examined the produced evaluation results on three levels: *i*) empirically, *ii*) analytically, *iii*) behaviourally. These experiments led to a better understanding of the behaviour of α -nDCG and the role of its parameter, compared to other common measures like s-recall and s-mrr (Chapter 8).

10.3 Future Work

Based on the work contained in this thesis, we identified several avenues for future research: we discuss them in the following.

Effective Sub-topic Modelling Techniques. In Section 6.2, we empirically evaluated a number of approaches for search result diversification. We highlighted that the approaches based on our integration framework can potentially improve the diversity effectiveness; however, the technique for estimating sub-topics is an impediment to the achievement of optimal performance. As can be observed from the results of our experiments, all three sub-topic modelling techniques fail to provide high quality evidence of sub-topics that correspond to users’ information needs. This issue limits diversification effectiveness that can be gained by integrating two ranking paradigms.

Alternative approaches to estimate sub-topics can be sought. For example, one may consider methods for supervised classification, which use a set of features to characterise documents into classes or sub-topics. These features are obtained from a sample set of documents (called a training set), of which classes are known and pre-defined for each query. Such a training set would then be employed by classification algorithms so as to learn how to classify *unknown* documents into known sub-topics. Apart from methods for classifying documents, alternative approaches may consider clustering techniques with *semantic* distances between documents. Examples include using a lexical ontology, e.g. WordNet¹ or DBpedia², to define semantic relations between words within documents [Lippincott and Passonneau, 2009].

Selective Result Diversification. In Section 6.2, we also evaluated two sample approaches based on inter-dependent document relevance paradigm, i.e. MMR and MPT. These diversification approaches are encoded by a similar ranking strategy, including tunable parameters to control a trade-off between promoting relevance and diversity in search results. Nevertheless, not all queries are equally ambiguous and their initial retrieval results (i.e. documents) have different distributions of sub-topics. As a result, different queries could benefit from different diversification strategies and hence parameter settings. Therefore, future work can be directed towards finding an effective approach to learn such a trade-off on a query-by-query basis and to set suitable parameters for each query. Santos et al. [2010] preliminarily investigated this approach using several query features. However, from the results of their study, we believe that substantial improvements in diversification performance are still possible by deploying more effective and sophisticated learning and feature selection techniques as well as other additional features, e.g. topic proportions of documents [Das et al., 2011].

Term and Topic Temporality for Document Diversification. Data collections such as web pages, news, and books consist of time-stamped documents covering many event-driven topics. Queries on these collections also contain temporal aspects associated with the topics within certain periods of time. We posit that *temporal profiling* [Whiting et al., 2011a] and *temporal locality* [Jin and Bestavros, 2000] can be

¹<http://wordnet.princeton.edu>

²<http://dbpedia.org>

exploited for effective time-based sub-topic modelling. Besides, they can be used to identify sub-topic popularity or intent-probability [Agrawal et al., 2009] within long-term or short-term periods such that the ordering of sub-topics can be taken into account with respect to this. Temporal profile is the time-based occurrence pattern of terms mined from a time-stamped collection of documents. Whiting et al. [2011a] exploited the temporal profile to improve retrieval effectiveness of pseudo-relevance feedback technique. Temporal locality is the probability that recently requested documents are likely to be requested again, and thus it reflects the temporally significant topics covered by such documents. Result diversification can therefore benefit from temporal features extracted from time-stamped queries and documents.

Set-Based Evaluation Measure for Diversified Search Results. In Section 7.4, we argued that sub-topic recall (s-recall) [Zhai et al., 2003] has three major drawbacks for diversity evaluation. First, s-recall is not a position-based metric. Second, s-recall does not take into account sub-topic redundancy. Third, s-recall cannot distinguish between retrieving relevant or non-relevant documents after complete sub-topic coverage is achieved. Consequently, any measure that incorporates s-recall possibly inherit its drawbacks. This is the case of, for example, D- and D#-measures [Sakai and Song, 2011]. With this respect, it would be interesting to develop a new evaluation measure by explicitly defining measures for diversity and redundancy separately. We believe that the diversity measure should be simple and based on set theory so that the diversity and coverage of sub-topic can be assessed. Moreover, the diversity measure should also consider rank positions of documents so as to attribute more importance to documents that are ranked on the top of the result list.

Examine the Intuitiveness of ERR-IA. In Section 7.5, we analysed the intuitiveness of α -nDCG in evaluating diversification performance. We discovered that arbitrarily setting α to 0.5 may turn α -nDCG to be counter-intuitive and not behave as anticipated by TREC evaluation guidelines. Recently, Intent-Aware Expected Reciprocal Rank (ERR-IA) [Chapelle et al., 2009] has been proposed as an alternative effectiveness measure for diversity evaluation. In addition, ERR-IA has been increasingly used by the IR community, e.g. it has been included as an official measure in the TREC Web

Diversity Track 2011. Nevertheless, the measure has not yet been thoroughly investigated in term of its validity to the task as well as its reliability in terms of discriminative power. It is undoubtedly an intriguing avenue of research to examine whether ERR-IA presents issue similar to those of α -nDCG or not.

Part VI

References and Appendices

References

- Eugene Agichtein, Eric Brill, and Susan Dumais. Improving Web Search Ranking by Incorporating User Behavior Information. In *Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '06, pages 19–26, Seattle, USA, 2006. 73, 88
- Rakesh Agrawal, Sreenivas Gollapudi, Alan Halverson, and Samuel Ieong. Diversifying Search Results. In *Proceedings of the 2nd ACM International Conference on Web Search and Data Mining*, WSDM '09, pages 5–14, Barcelona, Spain, 2009. 38, 85, 95, 97, 135, 145, 147, 197
- Ricardo Baeza-Yates and Berthier Ribeiro-Neto. *Modern Information Retrieval*. 1st. Addison-Wesley, 1999. 17, 33
- James E. Bartlett, Joe W. Kotrlik, and Chadwick C. Higgins. Organizational Research: Determining Appropriate Sample Size in Survey Research. *Information Technology, Learning and Performance*, 19(1):43–50, 2001. 120, 180
- Micheline Beaulieu and Susan Jones. Interactive Searching and Interface Issues in the Okapi Best Match Probabilistic Retrieval System. *Interacting with Computers*, 10(3):237 – 248, 1998. 27
- Yaniv Bernstein and Justin Zobel. Redundant Documents and Search Effectiveness. In *Proceedings of the 14th ACM international conference on Information and knowledge management*, CIKM '05, pages 736–743, Bremen, Germany, 2005. 25
- David Beymer and Daniel M. Russell. WebGazeAnalyzer: A System for Capturing and Analyzing Web Reading Behavior Using Eye Gaze. In *CHI '05 extended abstracts*

REFERENCES

- on Human factors in computing systems*, CHI EA '05, pages 1913–1916, Portland, USA, 2005. 48
- David M. Blei, Andrew Y. Ng, and Michael I. Jordan. Latent Dirichlet Allocation. *Journal of Machine Learning Research*, 3:993–1022, 2003. 85, 94, 96
- Dario Bonino, Alberto Ciaramella, and Fulvio Corno. Review of the State-of-the-Art in Patent Information and Forthcoming Evolutions in Intelligent Patent Informatics. *World Patent Information*, 32(1):30 – 38, 2010. 25
- Abraham Bookstein and Don R. Swanson. Probabilistic Models for Automatic Indexing. *Journal of the American Society for Information Science*, 25(5):312–316, 1974. 21, 86
- Pia Borlund. The IIR Evaluation Model: A Framework for Evaluation of Interactive Information Retrieval Systems. *Information Research*, 8(3), 2003a. 30, 63, 67
- Pia Borlund. The Concept of Relevance in IR. *Journal of the American Society for Information Science and Technology*, 54:913–925, 2003b. 29
- Pia Borlund and Peter Ingwersen. The Development of a Method for the Evaluation of Interactive Information Retrieval Systems. *Journal of Documentation*, 53:225–250, 1997. 30
- Bert Boyce. Beyond Topicality: A Two Stage View of Relevance and the Retrieval Process. *Information Processing & Management*, 18(3):105 – 109, 1982. 88
- Chris Buckley and Ellen M. Voorhees. Evaluating Evaluation Measure Stability. In *Proceedings of the 23rd annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '00, pages 33–40, Athens, Greece, 2000. 157, 175
- Georg Buscher, Andreas Dengel, and Ludger van Elst. Eye Movements as Implicit Relevance Feedback. In *CHI '08 extended abstracts on Human factors in computing systems*, CHI EA '08, pages 2991–2996, Florence, Italy, 2008. 48

REFERENCES

- Iain Campbell. Interactive Evaluation of the Ostensive Model Using a New Test Collection of Images with Multiple Relevance Assessments. *Information Retrieval*, 2 (1):89–114, 2000. [48](#), [54](#)
- Iain Campbell and Cornelis J. van Rijsbergen. The Ostensive Model of Developing Information Needs. In *Proceedings of the 2nd International Conference on Conceptions of Library and Information Science*, CoLIS '96, pages 251–268, Borås, Swedish, 1996. [43](#), [48](#), [53](#), [223](#)
- Jaime Carbonell and Jade Goldstein. The Use of MMR, Diversity-Based Reranking for Reordering Documents and Producing Summaries. In *Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '98, pages 335–336, Melbourne, Australia, 1998. [6](#), [84](#), [89](#), [90](#)
- Ben Carterette. An Analysis of NP-Completeness in Novelty and Diversity Ranking. In *Proceedings of the 2nd International Conference on Theory of Information Retrieval*, ICTIR '09, pages 89–106, Cambridge, UK, 2009. [147](#)
- Ben Carterette. System Effectiveness, User Models, and User Utility: A Conceptual Framework for Investigation. In *Proceedings of the 34th international ACM SIGIR conference on Research and development in Information*, SIGIR '11, pages 903–912, Beijing, China, 2011. [32](#), [142](#)
- Ben Carterette and Praveen Chandar. Probabilistic Models of Ranking Novel Documents for Faceted Topic Retrieval. In *Proceeding of the 18th ACM conference on Information and knowledge management*, CIKM '09, pages 1287–1296, Hong Kong, China, 2009. [84](#), [85](#), [94](#), [95](#), [97](#)
- Praveen Chandar and Ben Carterette. Analysis of Various Evaluation Measures for Diversity. In *Proceedings of the 33rd European Conference on Information Retrieval Workshop on Diversity in Document Retrieval*, DDR '11, pages 21–28, 2011. [162](#)
- Olivier Chapelle, Donald Metzler, Ya Zhang, and Pierre Grinspan. Expected Reciprocal Rank for Graded Relevance. In *Proceeding of the 18th ACM conference on Information and knowledge management*, CIKM '09, pages 621–630, Hong Kong, China, 2009. [35](#), [36](#), [38](#), [136](#), [141](#), [148](#), [178](#), [197](#)

REFERENCES

- Harr Chen and David R. Karger. Less is More: Probabilistic Models for Retrieving Fewer Relevant Documents. In *Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '06, pages 429–436, Seattle, Washington, USA, 2006. 37, 44, 88, 115, 140
- Joon Yeon Choeh and Hong Joo Lee. Mobile Push Personalization and User Experience. *AI Communications – Recommender Systems*, 21:185–193, 2008. 51
- Chales L.A. Clarke, Nick Craswell, and Ian Soboroff. Overview of the TREC 2009 Web Track. In *the 18th Text Retrieval Conference, 2009.*, 2009a. 23, 24, 113, 115, 131, 142, 159
- Chales L.A. Clarke, Nick Craswell, Ian Soboroff, and Gordon V. Cormack. Overview of the TREC 2010 Web Track. In *the 19th Text Retrieval Conference, 2010.*, 2010. 23, 24, 131, 142, 159
- Charles L.A. Clarke, Maheedhar Kolla, Gordon V. Cormack, Olga Vechtomova, Azin Ashkan, Stefan Büttcher, and Ian MacKinnon. Novelty and Diversity in Information Retrieval Evaluation. In *Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '08, pages 659–666, Singapore, 2008. 6, 7, 9, 25, 39, 44, 115, 141, 142, 143, 144, 145, 146, 154
- Charles L.A. Clarke, Maheedhar Kolla, and Olga Vechtomova. An Effectiveness Measure for Ambiguous and Underspecified Queries. In *Proceedings of the 2nd International Conference on Theory of Information Retrieval*, ICTIR '09, pages 188–199, Cambridge, UK, 2009b. 5, 141, 178
- C.L.A. Clarke, N. Craswell, I. Soboroff, and A. Ashkan. A Comparative Analysis of Cascade Measures for Novelty and Diversity. In *Proceedings of the 4th ACM international conference on Web search and data mining*, WSDM '11, pages 75–84. ACM, 2011. 145, 162
- Mark Claypool, Phong Le, Makoto Wased, and David Brown. Implicit Interest Indicators. In *Proceedings of the 6th international conference on Intelligent user interfaces*, IUI '01, pages 33–40, Santa Fe, USA, 2001. 48

REFERENCES

- Cyril W. Cleverdon. On the Inverse Relationship of Recall and Precision. *Journal of Documentation*, 28(3):195–201, 1972. 33
- Cyril W. Cleverdon. The Significance of the Cranfield Tests on Index Languages. In *Proceedings of the 14th annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '91, pages 3–12, Chicago, USA, 1991. ACM. 28
- Cyril W. Cleverdon, Jack Mills, and Michael Keen. Factors Determining the Performance of Indexing Systems. *Technical report, ASLIB Cranfield Project*, 1966. 26, 27
- William S. Cooper. The Suboptimality of Retrieval Rankings Based on Probability of Usefulness. Technical report, School of Library and Information Studies, University of Californiam Berkeley, 1976. 88
- Gordon V. Cormack, Maura R. Grossman, Bruce Hedin, and Douglas W. Oard. Overview of the TREC 2010 Legal Track. In *the 19th Text Retrieval Conference, 2010.*, 2010. 25
- Bruce Croft, Donald Metzler, and Trevor Strohman. *Search Engines: Information Retrieval in Practice*. Addison-Wesley, 1st edition, 2009. 19, 22
- Sujatha Das, Prasenjit Mitra, and C. Lee Giles. Learning to Rank Homepages For Researcher-Name Queries. In *Proceedings of the 1st International Workshop on Entity-Oriented Search*, EOS SIGIR '11, Beijing, China, 2011. 196
- Thomas Deselaers, Tobias Gass, Philippe Dreuw, and Hermann Ney. Jointly Optimising Relevance and Diversity in Image Retrieval. In *Proceeding of the ACM International Conference on Image and Video Retrieval*, CIVR '09, pages 1–8, Santorini, Greece, 2009. 50, 95, 97, 99, 110
- Michael Eisenberg and Carol Berry. Order Effects: A Atudy of the Possible Influence of Presentation Order on User Judgments of Document Relevance. *Journal of the American Society for Information Science and Technology*, 39(5):293–300, 2007. 88

REFERENCES

- Desmond Elliott, Frank Hopfgartner, Teerapong Leelanupab, Yashar Moshfeghi, and Joemon M. Jose. An Architecture for Life-long User Modelling. In *Proceedings of UMAP Workshop on Life-long User Modelling*, LLUM '09, pages 9–16, Trento, Italy, 2009. 15
- Marin Ferecatu and Hichem Sahbi. TELECOM ParisTech at ImageCLEFphoto 2008: Bi-Modal Text and Image Retrieval with Diversity Enhancement. In *Working Notes for the CLEF 2008 workshop*, Aarhus, Denmark, 2008. 99, 110
- Norbert Fuhr. A Probability Ranking Principle for Interactive Information Retrieval. *Information Retrieval*, 11:251–265, 2008. 84, 89
- Atsushi Fujii, Makoto Iwayama, and Noriko K. Overview of Patent Retrieval Task at NTCIR-4. In *In Proceedings of the Fourth NTCIR Workshop on Research in Information Access Technologies Information Retrieval, Question Answering and Summarization, 2004. of NTCIR-5 Workshop Meeting*, pages 6–9, 2004. 23
- Erich Gamma, Richard Helm, Ralph Johnson, and John Vlissides. *Design Patterns: Elements of Reusable Object-Oriented Software*. Addison-Wesley, 1995. 224
- Fredric C. Gey, Noriko Kando, and Carol Peters. Cross-Language Information Retrieval: the way ahead. *Information Processing & Management*, 41:415–431, 2005. 28
- Mark Girolami and Ata Kabán. On an Equivalence between PLSI and LDA. In *Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval*, SIGIR '03, pages 433–434, Toronto, Canada, 2003. 97
- William Goffman. On Relevance as a Measure. *Information Storage and Retrieval*, 2(3):201 – 203, 1964. 88
- William Goffman. An Indirect Method of Information Retrieval. *Information Storage and Retrieval*, 4(4):361 – 373, 1968. 87, 92
- Michael D. Gordon and Peter Lenk. A Utility Theoretic Examination of the Probability Ranking Principle in Information Retrieval. *Journal of the American Society for Information Science and Technology*, 42(10):703–714, 1991. 87, 89

REFERENCES

- Michael D. Gordon and Peter Lenk. When Is the Probability Ranking Principle Sub-optimal? *Journal of The American Society for Information Science and Technology*, 43:1–14, 1992. 83, 89
- Martin Halvey, P. Punitha, David Hannah, Robert Villa, Frank Hopfgartner, Anuj Goyal, and Joemon M. Jose. Diversity, Assortment, Dissimilarity, Variety: A Study of Diversity Measures Using Low Level Features for Video Retrieval. In *Proceedings of the 31th European Conference on IR Research on Advances in Information Retrieval*, ECIR '09, pages 126–137, Toulouse, France, 2009. 67, 97, 99
- David Hawking. Challenges in Enterprise Search. In *Proceedings of the 15th Australasian database conference - Volume 27*, ADC '04, pages 15–24, Dunedin, New Zealand, 2004. 23
- Marti A. Hearst. Next Generation Web Search: Setting Our Sites. *IEEE Data Engineering Bulletin*, 23(3):38–48, 2000. 45, 46
- Marti A. Hearst and Jan O. Pedersen. Reexamining the Cluster Hypothesis: Scatter/Gather on Retrieval Results. In *Proceedings of the 19th annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '96, pages 76–84, Zurich, Switzerland, 1996. 94
- David R. Heise. The semantic differential and attitude research. *Attitude measurement*, pages 235–253, 1970. 69
- Ivan Herman, Guy Melançon, and M. Scott Marshall. Graph Visualization and Navigation in Information Visualization. *IEEE Transactions on Visualization and Computer Graphics*, 6(1):24–43, 2000. 47
- William Hersh and Paul Over. TREC–8 Interactive Track Report. In *Proceedings of TREC–8*, pages 57–64, 2000. 43, 114
- Djoerd Hiemstra. *Using Language Models for Information Retrieval*. PhD thesis, Center for Telematics and Information Technology, University of Twente, 2011. 136
- Thomas Hofmann. Probabilistic Latent Semantic Indexing. In *Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '99, pages 50–57, Berkeley, USA, 1999. 85, 94, 96

REFERENCES

- Frank Hopfgartner, David Vallet, Martin Halvey, and Joemon Jose. Search Trails Using User Feedback to Improve Video Search. In *Proceedings of the 16th ACM international conference on Multimedia*, MM '08, pages 339–348, Vancouver, Canada, 2008. 67
- Jing Huang, S. Ravi Kumar, and Ramin Zabih. An Automatic Hierarchical Image Classification Scheme. In *Proceedings of the 6th ACM international conference on Multimedia*, MM '98, pages 219–228, Bristol, UK, 1998. 85
- Peter Ingwersen. *Information Retrieval Interaction*. Taylor Graham Publishing, 1992. 8, 41
- Peter Ingwersen. Cognitive perspectives of information retrieval interaction: Elements of a cognitive ir theory. *Journal of Documentation*, 52(1):3–50, 1996. 29, 42
- Peter Ingwersen and Kalervo Järvelin. *The Turn: Integration of Information Seeking and Retrieval in Context (The Information Retrieval Series)*. Springer-Verlag, 2005. 29, 42
- Anthony Jameson. *The Human-Computer Interaction Handbook: Evolving Technologies and Emerging Applications*, chapter Adaptive interfaces and agents, pages 433–458. L. Erlbaum Associates Inc., 2008. 50
- Kalervo Järvelin and Jaana Kekäläinen. Cumulated Gain-Based Evaluation of IR Techniques. *ACM Transactions on Information Systems*, 20(4):422–446, 2002. 35, 39, 115, 136, 141, 143
- Shudong Jin and Azer Bestavros. Sources and Characteristics of Web Temporal Locality. In *Proceedings of the 8th International Symposium on Modeling, Analysis and Simulation of Computer and Telecommunication Systems*, MASCOTS '00, San Francisco, USA, 2000. 196
- Hideo Joho, Leif A. Azzopardi, and Wim Vanderbauwhede. A Survey of Patent Users: An Analysis of Tasks, Behavior, Search Functionality and System Requirements. In *Proceeding of the third symposium on Information interaction in context*, volume Proceeding of the 3rd symposium on Information interaction in context of *IIIX '10*, pages 13–24, New Brunswick, USA, 2010. 25

REFERENCES

- Joemon M. Jose. *An Integrated Approach for Multimedia Information Retrieval*. PhD thesis, The Robert Gordon University, Aberdeen, 1998. 66
- Diane Kelly and Jaime Teevan. Implicit Feedback for Inferring User Preference: A Bibliography. *SIGIR Forum*, 37:18–28, 2003. 48
- Allen Kent, M. M. Berry, and J. W. Perry. Machine Literature Searching II. Problems in Indexing for Machine Searching. *American Documentation*, 5(1):22–25, 1954. 33
- Donald E. Knuth. *The Art of Computer Programming. Volume 2: Seminumerical Algorithms*. Addison-Wesley, 1969. 169
- Oren Kurland. *Inter-Document Similarities, Language Models, and Ad Hoc Information Retrieval (Doctoral dissertation)*. PhD thesis, Cornell University, 2006. 94
- Oren Kurland and Carmel Domshlak. A Rank-Aggregation Approach to Searching for Optimal Query-Specific Clusters. In *Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '08, pages 547–554, Singapore, 2008. 94
- Oren Kurland and Lillian Lee. Corpus Structure, Language Models, and Ad Hoc Information Retrieval. In *Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '04, pages 194–201, Sheffield, UK, 2004. 95, 97
- Victor Lavrenko and Bruce Croft, W. Relevance Based Language Models. In *Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '01, pages 120–127, New Orleans, USA, 2001. 94
- Teerapong Leelanupab and Joemon M. Jose. An Adaptive Browsing-Based Approach for Creating a Photographic Story. In *Proceedings of the 3th International Conference on Semantic and Digital Media Technologies: Semantic Multimedia*, SAMT '08, pages 196–197, Koblenz, Germany, 2008. 15

REFERENCES

- Teerapong Leelanupab, Yue Feng, Vassilios Stathopoulos, and Joemon M. Jose. A Simulated User Study of Image Browsing Using High-Level Classification. In *Proceedings of the 4th International Conference on Semantic and Digital Media Technologies: Semantic Multimedia*, SAMT '09, pages 3–15, Graz, Austria, 2009a. 14
- Teerapong Leelanupab, Martin Halvey, and Joemon M. Jose. Application and Evaluation of Multi-Dimensional Diversity. In *Proceedings of the Theseus/ImageCLEF workshop on visual information retrieval evaluation (Co-located with ECDL '09)*, pages 52–59, Corfu, Greece, 2009b. 14
- Teerapong Leelanupab, Frank Hopfgartner, and Joemon M. Jose. User Centred Evaluation of A Recommendation Based Image Browsing System. In *Proceedings of the 4th Indian International Conference on Artificial Intelligence*, IICAI '09, pages 558–573, Tumkur, India, 2009c. 14, 44
- Teerapong Leelanupab, Guido Zuccon, Anuj Goyal, Martin Halvey, P. Punitha, and Joemon M. Jose. University of Glasgow at ImageCLEFPhoto 2009: Optimising Similarity and Diversity in Image Retrieval. In *Multilingual Information Access Evaluation II. Multimedia Experiments*, volume 6242 of *CLEF' 09*, pages 133–141. Springer Berlin Heidelberg, 2010a. 14, 86
- Teerapong Leelanupab, Guido Zuccon, and Joemon M. Jose. Revisiting Sub-topic Retrieval in the ImageCLEF 2009 Photo Retrieval Task. In *ImageCLEF – experimental evaluation in image retrieval*, volume 32 of *The Information Retrieval Series*, chapter 15, pages 277–294. Springer Berlin Heidelberg, 2010b. 14, 86
- Teerapong Leelanupab, Guido Zuccon, and Joemon M. Jose. Technical Report: A Study of Ranking Paradigms and Their Integrations for Subtopic Retrieval. Technical report, Technical report, School of Computing Science, University of Glasgow, 2010c. 13
- Teerapong Leelanupab, Guido Zuccon, and Joemon M. Jose. When Two Is Better Than One: A Study of Ranking Paradigms and Their Integrations for Subtopic Retrieval. In *Information Retrieval Technology – the 6th Asia Information Retrieval Societies Conference*, AIRS '10, pages 162–172, Taipei, Taiwan, 2010d. 13, 86

REFERENCES

- Teerapong Leelanupab, Guido Zuccon, and Joemon M. Jose. A Query–Basis Approach to Parametrizing Novelty–Biased Cumulative Gain. In *Proceedings of the 3rd International Conference on Theory of Information Retrieval: Advances in Information Retrieval Theory*, ICTIR '09, Bertinoro, Italy, 2011. 13, 130
- Ting-Peng Liang, Hung-Jen Lai, and Yi-Cheng Ku. Personalized Content Recommendation and User Satisfaction: Theoretical Synthesis and Empirical Findings. *Journal of Management Information Systems*, 23:45–70, 2007. 50
- Thomas Lippincott and Rebecca Passonneau. Semantic Clustering for a Functional Text Classification Task. In *Proceedings of the 10th International Conference on Computational Linguistics and Intelligent Text Processing*, CICLing '09, Mexico City, Mexico, 2009. 196
- Xiaoyong Liu and W. Bruce Croft. Evaluating Text Representations for Retrieval of the Best Group of Documents. In *Proceedings of the 30th European Conference on IR Research on Advances in Information Retrieval*, ECIR '08, pages 454–462, Glasgow, UK, 2008. 94
- Hans Peter Luhn. A Statistical Approach to Mechanized Encoding and Searching of Literary Information. *IBM Journal of Research and Development*, 1:309–317, 1957. 17
- Hans Peter Luhn. The Automatic Creation of Literature Abstracts. *IBM Journal of Research and Development*, 2:159–165, 1958. 17
- J. B. MacQueen. Some Methods of Classification and Analysis of Multivariate Observations. In *Proceedings of 5th Berkeley Symposium on Mathematical Statistics and Probability*, pages 281–297, 1967. 84, 94, 95
- Paul P. Maglio, Rob Barrett, Christopher S. Campbell, and Ted Selker. SUITOR: An Attentive Information System. In *Proceedings of the 5th international conference on Intelligent user interfaces*, IUI '00, pages 169–176, New Orleans, USA, 2000. ACM. 48
- Christopher D. Manning, Prabhakar Raghavan, and Hinrich Schtze. *Introduction to Information Retrieval*. Cambridge University Press, 2008. 16, 33

REFERENCES

- Gary Marchionini. Exploratory Search: From Finding to Understanding. *Communications of the ACM*, 49(4):41–46, 2006. 41, 45, 46
- M. E. Maron and J. L. Kuhns. On Relevance, Probabilistic Indexing and Information Retrieval. *Journal of the ACM*, 7:216–244, 1960. 86, 133
- George A. Miller. The Magical Number Seven, Plus or Minus Two: Some Limits on Our Capacity for Processing Information. *The Psychological Review*, 63:81–97, 1956. 58
- Alistair Moffat and Justin Zobel. Rank-Biased Precision for Measurement of Retrieval Effectiveness. *ACM Transactions on Information Systems*, 27:2:1–2:27, 2008. 136, 141
- Allen Newell and Herbert Alexander Simon. *Human Problem Solving*. Prentice-Hall, 1972. 5
- Noel E. O’Connor. Multimedia Signal Processing: An Overview for Content-based Information Retrieval. Summer School on Multimedia Semantics, 2009. 249, 252
- Iadh Ounis, Christina Lioma, Craig Macdonald, and Vassilis Plachouras. Research Directions in Terrier: a Search Engine for Advanced Retrieval on the Web. *Novatica/UPGRADE Special Issue on Web Information Access, Ricardo Baeza-Yates et al. (Eds), Invited Paper*, 2007. 66, 223
- Paul Over. TREC–6 Interactive Track Report. In *Proceedings of TREC–6*, pages 73–82, 1998. 43, 114
- Paul Over. TREC–7 Interactive Track Report. In *Proceedings of TREC–7*, pages 65–72, 1999. 43, 114
- Paul Over. The TREC Interactive Track: An Annotated Bibliography. *Information Processing and Management*, 37(3):369 – 381, 2001. 24, 82, 137
- Monica Lestari Paramita, Mark Sanderson, and Paul Clough. Developing a Test Collection to Support Diversity Analysis. In *Proceedings of Redundancy, Diversity, and IDR workshop held at SIGIR ’09*, pages 39–45, 2009. 112

REFERENCES

- Michael J. Pazzani and Daniel Billsus. Content-Based Recommender Systems. *The Adaptive Web: Lecture Notes in Computer Science*, pages 325–341, 2007. 51
- Peter Pirolli, Patricia Schank, Marti Hearst, and Christine Diehl. Scatter/Gather Browsing Communicates the Topic Structure of a Very Large Text Collection. In *Proceedings of the SIGCHI conference on Human factors in computing systems: common ground*, CHI '96, pages 213–220, Vancouver, Canada, 1996. 46
- Jay M. Ponte and W. Bruce Croft. A Language Modeling Approach to Information Retrieval. In *Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '98, pages 275–281, 1998. 136
- Filip Radlinski and Susan Dumais. Improving Personalized Web Search Using Result Diversification. In *Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '06, pages 691–692, Seattle, Washington, USA, 2006. 25
- Filip Radlinski, Paul N. Bennett, Ben Carterette, and Thorsten Joachims. Redundancy, Diversity and Interdependent Document Relevance. *ACM SIGIR Forum*, 43:46–52, 2009. 7
- Vijay Raghavan, Peter Bollmann, and Gwang S. Jung. A Critical Investigation of Recall and Precision as Measures of Retrieval System Performance. *ACM Transactions on Information Systems*, 7:205–229, 1989. 33
- Yong Man Ro, Munchurl Kim, Ho Kyung Kang, B. S. Manjunath, and Jinwoong Kim. MPEG-7 Homogeneous Texture Descriptor. *ETRI*, 23:41–51, 2001. 255, 256
- Phoebe M. Roberts, Aaron M. Cohen, and William R. Hersh. Tasks, Topics and Relevance Judging for the TREC Genomics Track: Five Years of Experience Evaluating Biomedical Text Information Retrieval Systems. *Information Retrieval*, 12:81–97, 2009. 23
- Stephen E. Robertson. The Probability Ranking Principle in IR. *Journal of Documentation*, 33(4):294–304, 1977. 20, 83, 87, 105, 136

REFERENCES

- Stephen E. Robertson and Nicholas J. Belkin. Ranking in Principle. *Journal of Documentation*, 34(2):93–100, 1978. 136
- Stephen E. Robertson and Micheline M. Hancock-Beaulieu. On the Evaluation of IR Systems. *Information Processing & Management*, 28(4):457 – 466, 1992. 29
- Stephen E. Robertson and Ian Soboroff. The TREC–10 Filtering Track Final Report. *Proceeding of the Tenth Text REtrieval Conference (TREC-10)*, pages 26–37, 2002. 23
- Stephen E. Robertson and Karen Spärck-Jones. Relevance Weighting of Search Terms. *Journal of the American Society for Information Science*, 27(3):129–146, 1976. 20, 21, 86, 87, 136
- Stephen E. Robertson and Stephen G. Walker. Some Simple Effective Approximations to the 2-Poisson Model for Probabilistic Weighted Retrieval. In *Proceedings of the 17th annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '94, pages 232–241, Dublin, Ireland, 1994. 21
- Stephen E. Robertson, Stephen G. Walker, Karen Spärck-Jones, Micheline M. Hancock-Beaulieu, and Mike Gatford. Okapi at TREC–3. In *Proceedings of the Third Text REtrieval Conference (TREC 1994)*, Gaithersburg, USA, 1994. 21, 86
- Stephen E. Robertson, Stephen G. Walker, Micheline M. Hancock-Beaulieu, M. Gatford, and A. Payne. Okapi at TREC–4. In *Proceedings of TREC–4*, 1996. 116
- Stefan M Rüger. Keynote Talk: More Than a Thousand Words. In *Proceedings of the 4th International Conference on Semantic and Digital Media Technologies: Semantic Multimedia*, SAMT '09, pages 2–2, Graz, Austria, 2009. 57
- Yong Rui, Thomas S. Huang, and Sharad Mehrotra. Relevance Feedback Techniques in Interactive Content-Based Image Retrieval. In *Proceedings of the Sixth Conference on Storage and Retrieval for Image and Video Databases*, volume 3312 of *SPIE*, pages 25–36, San Jose, California, USA, 1998a. 47
- Yong Rui, Thomas S. Huang, Michael Ortega, and Sharad Mehrotra. Relevance Feedback: A Power Tool for Interactive Content-Based Image Retrieval. *Circuits and Systems for Video Technology, IEEE Transactions on*, 8(5):644–655, 1998b. 47

REFERENCES

- Ian Ruthven. Integration Approaches to Relevance. In *New Directions in Cognitive Information Retrieval*, volume 19 of *The Information Retrieval Series*, chapter 4, pages 61–80. Springer Berlin Heidelberg, 2005. 29
- Tetsuya Sakai. Evaluating Evaluation Metrics Based on the Bootstrap. In *Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '06, pages 525–532, Seattle, Washington, USA, 2006. 157, 172, 173, 175
- Tetsuya Sakai and Stephen Robertson. Modelling a User Population for Designing Information Retrieval Metrics. In *Proceedings of the 2nd International Workshop on Evaluating Information Access*, EVIA '08, pages 30–41, 2008. 143
- Tetsuya Sakai and Ruihua Song. Evaluating Diversified Search Results Using Per-intent Graded Relevance. In *Proceedings of the 34th annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '11, Beijing, China, 2011. 132, 145, 166, 197
- Tetsuya Sakai, Nick Craswell, Ruihua Song, Stephen Robertson, Zhicheng Dou, and Chin Yew Lin. Simple Evaluation Metrics for Diversified Search Results. In *Proceedings of the 3rd International Workshop on Evaluating Information Access*, EVIA '10, pages 42–50, 2010. 38
- Phillipe Salembier and Thomas Sikora. *Introduction to MPEG-7: Multimedia Content Description Interface*. John Wiley & Sons, 2002. 223, 251, 252, 253, 254, 255, 256
- Gerard Salton. *The SMART Retrieval System—Experiments in Automatic Document Processing*. Prentice-Hall, 1971. 33
- Gerard Salton and Chris Buckley. Improving Retrieval Performance by Relevance Feedback. *Readings in information retrieval*, pages 355–364, 1997. 47
- Gerard Salton and Michael J. McGill. *Introduction to Modern Information Retrieval*. McGraw-Hill, 1986. 5, 18
- Gerard Salton, Andrew Wong, and Chung Shu Yang. A Vector Space Model for Automatic Indexing. *Communications of the ACM*, 18:613–620, 1975. 18

REFERENCES

- Gerard Salton, Edward A. Fox, and Harry Wu. Extended Boolean Information Retrieval. *Communications of the ACM*, 26:1022–1036, 1983. 17
- Mark Sanderson. Ambiguous Queries: Test Collections Need More Sense. In *Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '08, pages 499–506, Singapore, 2008. 24
- Mark Sanderson and Hideo Joho. Forming Test Collections with No System Pooling. In *Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '04, pages 33–40, Sheffield, United Kingdom, 2004. 29
- Mark Sanderson, Monica Lestari Paramita, Paul Clough, and Evangelos Kanoulas. Do User Preferences and Evaluation Measures Line Up? In *Proceeding of the 33rd international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '10, pages 555–562, Geneva, Switzerland, 2010. 148
- Rodrygo L. T. Santos, Craig Macdonald, and Iadh Ounis. On the Suitability of Diversity Metrics for Learning-to-Rank for Diversity. In *Proceedings of the 34th international ACM SIGIR conference on Research and development in Information*, SIGIR '11, Beijing, China, 2011a. 132
- Rodrygo L.T. Santos, Craig Macdonald, and Iadh Ounis. Selectively Diversifying Web Search Results. In *Proceedings of the 19th ACM international conference on Information and knowledge management*, CIKM '10, pages 1179–1188, Toronto, Canada, 2010. 85, 86, 196
- Rodrygo L.T. Santos, Craig Macdonald, and Iadh Ounis. Intent-Aware Search Result Diversification. In *Proceedings of the 34th international ACM SIGIR conference on Research and development in Information Retrieval*, SIGIR '11, pages 595–604, Beijing, China, 2011b. 25
- J. Ben Schafer, Dan Frankowski, Jon Herlocker, and Shilad Sen. Collaborative Filtering Recommender Systems. *The Adaptive Web: Lecture Notes in Computer Science*, pages 291–324, 2007. 51

REFERENCES

- Young Woo Seo and Byoung Tak Zhang. Learning User's Preferences by Analyzing Web-Browsing Behaviors. In *Proceedings of the fourth international conference on Autonomous agents*, AGENTS '00, pages 381–387, Barcelona, Spain, 2000. 48
- William R. Shadish, Thomas D. Cook, and Donald T. Campbell. *Experimental and Quasi-Experimental Designs for Generalized Causal Inference*. Houghton Mifflin, 2nd edition, 2001. 30
- Herbert A. Simon. The Structure of Ill Structured Problems. *Artificial Intelligence*, 4 (3-4):181 – 201, 1973. 5
- Arnold W. M. Smeulders, Marcel Worring, Simone Santini, Amarnath Gupta, and Ramesh Jain. Content-Based Image Retrieval at the End of the Early Years. *IEEE Transactions on Pattern Analysis Machine Intelligence*, 22(12):1349–1380, 2000. 46
- Ruihua Song, Zhenxiao Luo, Jian-Yun Nie, Yong Yu, and Hsiao-Wuen Hon. Identification of Ambiguous Queries in Web Search. *Information Processing Management*, 45:216–229, 2009. 25
- Karen Spärck-Jones and Cornelis J. van Rijsbergen. Report on the Need for and Provision of an “Ideal” Information Retrieval Test Collection. Technical Report 5266, Computer Laboratory, University of Cambridge, 1975. 28
- Karen Spärck-Jones, Stephen E. Robertson, and Mark Sanderson. Ambiguous Requests: Implications for Retrieval Tests, Systems and Theories. *ACM SIGIR Forum*, 41:8–17, 2007. 24
- Keith H. Stirling. *The Effect of Document Ranking on Retrieval System Performance: A Search for an Optimal Ranking Rule*. PhD thesis, University of California, 1977. 88, 133, 135
- Keith H. Stirling. On the Limitations of Document Ranking Algorithms in Information Retrieval. In *Proceedings of the 4th annual international ACM SIGIR conference on Information storage and retrieval: theoretical issues in information retrieval*, SIGIR '81, pages 63–65, Oakland, California, 1981. 83

REFERENCES

- Kar Yan Tam and Shuk Ying Ho. Understanding the Impact of Web Personalization on User Information Precessing and Decision Outcomes. *MIS Quarterly*, 30(4): 865–890, 2006. 50
- A H M Ter Hofstede, H A Proper, and T P Van Der Weide. Query Formulation as an Information Retrieval Problem. *The Computer Journal*, 39(4):255–274, 1996. 42
- Anastasios Tombros, Robert Villa, and Cornelis J. van Rijsbergen. The Effectiveness of Query-Specific Hierarchic Clustering in Information Retrieval. *Information Processing & Management*, 38:559–582, 2002. 94
- Jana Urban, Joemon M. Jose, and Cornelis J. van Rijsbergen. An Adaptive Technique for Content-Based Image Retrieval. *Multimedia Tools and Applications*, 31(1):1–28, 2006. 8, 43, 48, 49, 52, 67, 222
- Thierry Urruty, Frank Hopfgartner, David Hannah, Desmond Elliott, and Joemon M. Jose. Supporting Aspect-Based Video Browsing - Analysis of a User Study. In *Proceeding of the ACM International Conference on Image and Video Retrieval, CIVR '09*, pages 1–8, Santorini, Greece, 2009. 58, 99
- Reinier H. van Leuken, Lluís Garcia, Ximena Olivares, and Roelof van Zwol. Visual Diversification of Image Search Results. In *Proceedings of the 18th international conference on World wide web, WWW '09*, pages 341–341, Madrid, Spain, 2009. 57
- Cornelis J. van Rijsbergen. *Information Retrieval*. Butterworth, 2nd edition, 1979. 19, 26, 33, 94, 116
- Marie-Luce Viaud, Jérôme Thièvre, Hervé Goëau, Agnes Saulnier, and Olivier Buisson. Interactive Components for Visual Exploration of Multimedia Archives. In *Proceedings of the 2008 international conference on Content-based image and video retrieval, CIVR '08*, pages 609–616, Niagara Falls, Canada, 2008. 47
- Robert Villa, Nicholas Gildea, and Joemon M. Jose. FacetBrowser: A User Interface for Complex Search Tasks. In *Proceedings of the 16th ACM international conference on Multimedia, MM '08*, pages 489–498, Vancouver, Canada, 2008. 67

REFERENCES

- Robert Villa, Iván Cantador, Hideo Joho, and Joemon M. Jose. An Aspectual Interface for Supporting Complex Search Tasks. In *Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '09, pages 379–386, Boston, USA, 2009. 46, 52
- Ellen M. Voorhees. TREC–8 Question Answering Track Report. In *Proceedings of the 8th Text Retrieval Conference*, pages 77–82, 1999. 140
- Ellen M. Voorhees. TREC: Improving Information Access through Evaluation. *Bulletin of the American Society for Information Science and Technology*, 32(1):16–21, 2005. 27
- Ellen M. Voorhees and Chris Buckley. The Effect of Topic Set Size on Retrieval Experiment Error. In *Proceedings of the 25th annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '02, pages 316–323, Tampere, Finland, 2002. 157, 175, 176
- Ellen M. Voorhees and Donna K. Harman. Proceedings of Text REtrieval Conference (TREC1–9). In *NIST Special Publications*, 2001. 82, 87
- Ellen M. Voorhees and Donna K. Harman. *TREC: Experiment and Evaluation in Information Retrieval*. MIT Press, 2005. 24, 28, 34, 67, 120, 180
- Jun Wang and Jianhan Zhu. Portfolio Theory of Information Retrieval. In *Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '09, pages 115–122, Boston, USA, 2009. 37, 84, 89, 91, 116, 117, 140
- Ryen W. White and Resa A. Roth. Exploratory Search: Beyond the Query-Response Paradigm. *Synthesis Lectures on Information Concepts, Retrieval, and Services*, 1: 1–98, 2009. 5, 8, 41, 46
- Ryen W. White, Bill Kules, Steven M. Drucker, and Schraefel M.C. Supporting Exploratory Search. *Communications of the ACM*, 49(4):36–39, 2006. 42

REFERENCES

- Ryen W. White, Gary Marchionini, and Gheorghe Muresan. Evaluating Exploratory Search Systems: Introduction to Special Topic Issue of Information Processing and Management. *Information Processing & Management*, 44(2):433–436, 2008. 24, 42, 45
- Stewart Whiting, Yashar Moshfeghi, and Joemon M. Jose. Exploring Term Temporality for Pseudo-Relevance Feedback. In *Proceedings of the 34th international ACM SIGIR conference on Research and development in Information Retrieval*, SIGIR '11, Beijing, China, 2011a. 196, 197
- Stewart Whiting, Jesus Rodriguez Perez, Guido Zuccon, Teerapong Leelanupab, and Joemon M. Jose. University of Glasgow (qirdcsuog) at TREC Crowdsourcing 2001: TurkRank – Network–Based Worker Ranking in Crowdsourcing. In *Proceedings of Text REtrieval Conference*. NIST, 2011b. 14
- Yiming Yang, Abhimanyu Lad, Ni Lao, Abhay Harpale, Bryan Kisiel, and Monica Rogati. Utility-Based Information Distillation Over Temporally Sequenced Documents. In *Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '07, pages 31–38, Amsterdam, Netherlands, 2007. 23
- Emine Yilmaz, Javed A. Aslam, and Stephen Robertson. A New Rank Correlation Coefficient for Information Retrieval. In *Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '08, pages 587–594, 2008. 161
- Emine Yilmaz, Milad Shokouhi, Nick Craswell, and Stephen Robertson. Expected Browsing Utility for Web Search Evaluation. In *Proceedings of the 19th ACM international conference on Information and knowledge management*, CIKM '10, pages 1561–1564, Toronto, Canada, 2010. 148
- Cheng Xiang Zhai, William W. Cohen, and John Lafferty. Beyond Independent Relevance: Methods and Evaluation Metrics for Subtopic Retrieval. In *Proceedings of the 26th annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '03, pages 10–17, Toronto, Canada, 2003. 6, 24, 25, 37, 82, 88, 89, 114, 115, 139, 197

REFERENCES

- ChengXiang Zhai and John Lafferty. A Risk Minimization Framework for Information Retrieval. *Information Processing & Management*, 42(1):31 – 55, 2006. 44, 88
- Yuye Zhang, Laurence A. Park, and Alistair Moffat. Click-Based Evidence for Decaying Weight Distributions in Search Effectiveness Metrics. *Information Retrieval*, 13:46–69, 2010. 148
- Zhong Qiu Zhao and Herve Glotin. Diversifying Image Retrieval by Affinity Propagation Clustering on Visual Manifolds. *IEEE MultiMedia*, 99(1), 2009. 95, 97, 99, 110
- Cai Nicolas Ziegler, Sean M. McNee, Joseph A. Konstan, and Georg Lausen. Improving Recommendation Lists through Topic Diversification. In *Proceedings of the 14th international conference on World Wide Web*, WWW '05, pages 22–32. ACM, 2005. ISBN 1-59593-046-9. 51
- Justin Zobel and Alistair Moffat. Inverted Files for Text Search Engines. *ACM Computing Surveys*, 38, 2006. 18
- Guido Zuccon and Leif Azzopardi. Using the Quantum Probability Ranking Principle to Rank Interdependent Documents. In *Proceedings of the 32nd European Conference on IR Research on Advances in Information Retrieval*, ECIR '10, pages 357–369, Milton Keynes, UK, 2010. 37, 140
- Guido Zuccon, Leif Azzopardi, and Cornelis J. van Rijsbergen. The Quantum Probability Ranking Principle for Information Retrieval. In *Proceedings of the 2nd International Conference on Theory of Information Retrieval: Advances in Information Retrieval Theory*, ICTIR '09, pages 232–240, Cambridge, UK, 2009a. 84, 89, 92
- Guido Zuccon, Teerapong Leelanupab, Anuj Goyal, Martin Halvey, P. Punitha, and Joemon M. Jose. The University of Glasgow at ImageClefPhoto 2009. In *Working Notes of Image Retrieval in CLEF 2009*, ImageCLEF '09, Corfu, Greece, 2009b. 14
- Guido Zuccon, Leif Azzopardi, and Cornelis J. van Rijsbergen. The Interactive PRP for Diversifying Document Rankings. In *Proceedings of the 34th international ACM SIGIR conference on Research and development in Information*, SIGIR '11, pages 1227–1228, Beijing, China, 2011a. 84

REFERENCES

Guido Zuccon, Teerapong Leelanupab, Stewart Whiting, Joemon M. Jose, and Leif Azzopardi. Crowdsourcing Interactions: A Proposal for Capturing User Interactions through Crowdsourcing. In *Proceedings of WSDM Workshop on Crowdsourcing for Search and Data Mining*, CSDM '11, Hong Kong, China, 2011b. 15

Guido Zuccon, Teerapong Leelanupab, Stewart Whiting, Emine Yilmaz, Joemon M. Jose, and Leif Azzopardi. Crowdsourcing Interactions: Capturing Query Sessions through Crowdsourcing. In *Proceedings of ECIR Workshop on Information Retrieval Over Query Sessions*, SIR '11, Dublin, Ireland, 2011c. 15

Appendix A

Architecture and Implementation of *Ostensive Browser Plus*

The *Ostensive Browser Plus* (OBP) system is purely implemented in Java and therefore is platform independent; it has been tested on three different platforms, i.e., Microsoft Windows, Linux, and Mac OS X. To run the system, a machine with at least 1GB of RAM and a single-core processor of 1.8 GHz or above is recommended. The system is pre-configured for a wide-screen display capable of a 16:10 aspect ratio resolution, e.g., 1280×800, 1440×900. Although the system is compatible with other screen resolutions, it is required to adjust the configuration settings in a properties file to accommodate for bigger or smaller screens.

The OBP system was built by refining and improving the original Ostensive Browser (OB) system developed by [Urban et al. \[2006\]](#). Over 20 new features have been added to the original system. Some of the key features used in a user study are presented in Chapter 3 are:

- diverse recommendations based on implicit feedback of user browsing interactions
- aspectual browsing interface
- animated visualisation of presentation
- client-server application
- full drag and drop support

The OBP consists of 1000 Java classes, of which more than half implement the interface. The visual features used in OBP are adapted from the CoPhIR¹ collection developed as part of the SAPIR² project. The interface is based on Java Swing and Abstract Windows Toolkit (AWT), and the animation feature for presentation is implemented by JOGL³, the Java Binding for the OpenGL API. The system is integrated with the Terrier retrieval toolkit⁴ [Ounis et al., 2007] for indexing text associated with images. For content-based image retrieval, the system supports various image distance measures depending on the visual features that are used. Note that similarity measures between visual features are implemented according to the MPEG-7 standard [Salem-bier and Sikora, 2002] and described in Appendix C. The system is organised into three main packages:

Server-side packages

- *ostensive.server.data* for the data representation including classes for the available document types, a class representing a collection, etc.
- *ostensive.server.feature* for the visual feature extractors.
- *ostensive.server.irmodel* for all IR related classes. The most important classes are the DocumentFactory, which manages the various document indices, and the RetrievalEngine, which incorporates ostensive relevance profiles [Campbell and van Rijsbergen, 1996] and links Terrier libraries with the system for retrieval operations.
- *ostensive.server.recommendation* for the diverse recommendations including a class for the graph-based representation of implicit feedback, a class for visual-based clustering of images, etc.

¹<http://cophir.isti.cnr.it/>

²<http://www.sapir.eu/>

³<http://jogamp.org/jogl/www/>

⁴<http://ir.dcs.gla.ac.uk/terrier/>

Helper (Mediator) packages

- *ostensive.common* for the communication class functioning as an interface to a server side. This class is implemented according to the Proxy Design pattern [[Gamma et al., 1995](#)]. It controls object access and establishes connections between client and server.

Client-side packages

- *ostensive.client.gui* for all interface-related objects.

Appendix B

Diverse Recommendations Through Image Browsing: Experimental Documents

This appendix presents the experimental documents described in the Chapter 3. These include:

B.1: Information Sheet

B.2: Consent Form

B.3: Task Descriptions

B.4: Entry Questionnaire

B.5: Post-Search Questionnaire for Baseline System

B.6: Post-Search Questionnaire for Recommender System

B.7: Exit Questionnaire

INFORMATION SHEET

Project: A Study of Diverse Recommendations to Support Exploratory Search in Image Browsing System

Researcher: Teerapong Leelanupab



**UNIVERSITY
of
GLASGOW**

You are invited to take part in a research study. Before you decide to do so, it is important for you to understand why the research is being done and what it will involve. Please take time to read the following information carefully. Ask me if anything is not clear or if you would like more information.

The objective of this experiment is to evaluate and compare the relative effectiveness of two different image search systems. Both systems have two default features: *i*) typical search feature, i.e. search by query, and *ii*) a browsing feature, which allows you to explore an image collection by selection of retrieved images. In addition to the default features, only one system contains interactive recommendation that instantly provides you with additional images based on the trail of your browsing interactions. The value of search systems cannot be evaluated unless we ask the people who are likely to using them. This is why your cooperation is needed to join our experiments. Please remember that it is the systems, not you, that are being evaluated.

It is up to you to decide whether or not to take part. If you decide to take part you will be given this information sheet to keep and asked to sign a consent form. You are free to withdraw at any time without giving a reason. You also have the right to withdraw retrospectively any consent given, and to require destroying any data gathered on you.

The experiment will last around two and a half hours and you will receive compensation of £15 upon completion. You will carry out four image search tasks using two different search systems. You will be given a chance to learn how to use all the two systems before we begin. At this time you will also be asked to complete an introductory questionnaire. You will perform three tasks in total. There is a time limit for each task, which takes 20 minutes. After completing each task you will be asked to fill in a questionnaire about your experience during the search and all of your interactions (e.g., mouse clicks and key presses) will also be logged. You are encouraged to comment on each interface as you use it, which I will take notes on. Please ask questions if you need and please let me know when you are finished with the task. Finally, after completing all tasks, you will be asked some questions about the tasks, your search strategy and the systems. Remember, you can opt out at any time during the experiment. You will still be rewarded for your effort depending on the number of tasks completed.

All information collected about you during the course of this study will be kept strictly confidential. You will be identified by an ID number and your information that contains name and contact details will be removed so that you cannot be recognised from it. Data will be only used for this study, and then destroyed. The results of this study may be used for some PhD research. You will not be identified in any report or publication that arises from this work.

This study is being funded by the Royal Thai Government PhD. Scholarship and European K-Space projects at the Department of Computer Science, University of Glasgow. This project has been reviewed by the Faculty of Information and Mathematical Sciences Ethics Committee.

For further information about this study please contact

Teerapong Leelanupab
Department of Computing Science, University of Glasgow
18 Lilybank Gardens
Glasgow, G12 8QQ
Email: kimm@dcs.gla.ac.uk
Tel.: 0141 330 1641



CONSENT FORM



Project: A Study of Diverse Recommendations to Support
Exploratory Search in Image Browsing System

Researcher: Teerapong Leelanupab

Please tick box

1. I confirm I have read and understand the information sheet for the above study and have had the opportunity to ask questions. ☐
2. I understand that my permission is voluntary and that I am free to withdraw at any time, without giving any reason, without my legal rights being affected. ☐
3. I agree to take part in the above study. ☐
4. I would like to receive a summary sheet of the experimental findings ☐

If you wish a summary, please leave an email address _____

Name of Participant Date Signature

Researcher Date Signature

TASK DESCRIPTION

Project: A Study of Diverse Recommendations to Support Exploratory Search in Image Browsing System

Researcher: Teerapong Leelanupab



Task A: *Wild Living Creatures*

Task Scenario:

Imagine you are a graphic designer of an activist organization for wildlife rehabilitation. Your task is to prepare an image presentation on various subjects of the Wildlife Conservation (WLC). The presentation is aimed at calling general awareness for endangered species and preservation of their habitats. You want to create a short presentation about the variety of wild living creatures.

Your task is to find as many relevant images as possible and save them for presentation. Each image must contain at least one different aspect (sub-topic) that complements the task described above. If one image covers several such aspects, then you need *not* to save other images that repeat those aspects. The examples of what constitutes aspect diversity are given below, but not limited to them. You are free to think of any other aspect that suits the task.

Indicative Request:

Your task is to find, using the provided system, relevant images showing different species of wild animals. The images you have to find should cover at least the following aspects:

- Terrestrial animals,
- Aquatic animals,
- Birds, etc.

TASK DESCRIPTION

Project: A Study of Diverse Recommendations to Support Exploratory Search in Image Browsing System

Researcher: Teerapong Leelanupab



UNIVERSITY
of
GLASGOW

Task B: *Man-made Vehicles*

Task Description

Task Scenario:

Imagine you are the decorator of the transportation museum. You want to create a short educational presentation about the variety of vehicles humans have built to operate in various situations, such as transportation, conveyance, or sport competition.

Your task is to find as many relevant images as possible and save them for presentation. Each image must contain at least one different aspect (sub-topic) that complements the task described above. If one image covers several such aspects, then you need *not* to save other images that repeat those aspects. The examples of what constitutes aspect diversity are given below, but not limited to them. You are free to think of any other aspect that suits the task.

Indicative Request:

Your task is to find, using the provided system, relevant images showing different types of vehicles. The images you have to find should cover at least the following aspects:

- Car,
- Train,
- Ship, etc.

TASK DESCRIPTION

Project: A Study of Diverse Recommendations to Support Exploratory Search in Image Browsing System

Researcher: Teerapong Leelanupab



UNIVERSITY
of
GLASGOW

Task C: *Marine Ecology*

Task Description

Task Scenario:

Imagine you are a student assigned to find information about the importance of the marine ecological system. You decide to make a presentation about natural water resources that are currently contaminated by pollution or harmed by human activities. You want to create a short presentation about a variety of water resources in order to convince the public to pay attention to their importance.

Your task is to find as many relevant images as possible and save them for presentation. Each image must contain at least one different aspect (sub-topic) that complements the task described above. If one image covers several such aspects, then you need *not* to save other images that repeat those aspects. The examples of what constitutes aspect diversity are given below, but not limited to them. You are free to think of any other aspect that suits the task.

Indicative Request:

Your task is to find, using the provided system, relevant images showing different natural water resources. The images you have to find should cover at least the following aspects:

- Headspring,
- Estuary,
- River, etc.

TASK DESCRIPTION

Project: A Study of Diverse Recommendations to Support Exploratory Search in Image Browsing System

Researcher: Teerapong Leelanupab



Task D: *Beautiful British Scenery*

Task Description

Task Scenario:

Imagine you are an officer of the British tourism agency. You are responsible for marketing Britain worldwide and promoting British tourism. At the time, your team would like to promote tourism in the countryside of Britain. You decide to make the multimedia presentation to show spectacular views of scenery in UK so as to attract foreign visitors.

Your task is to find as many relevant images as possible and save them for presentation. Each image must contain at least one different aspect (sub-topic) that complements the task described above. If one image covers several such aspects, then you need *not* to save other images that repeat those aspects. The examples of what constitutes aspect diversity are given below, but not limited to them. You are free to think of any other aspect that suits the task.

Indicative Request:

Your task is to find, using the provided system, relevant images showing the scenes of different attractive places in rural areas of UK for visitors. The images you have to find should cover at least the following aspects:

- Cliff,
- Mountain,
- Castle, etc.

ENTRY QUESTIONNAIRE

This questionnaire will provide us with background information that will help us analyse the answers you give in later stages of this experiment. You are not obliged to answer a question, if you feel it is too personal.



**UNIVERSITY
of
GLASGOW**

User ID:

Please place a TICK ☒ in the square that best matches your opinion.

Part 1: PERSONAL DETAILS

This information is kept completely confidential and no information is stored on computer media that could identify you as a person.

1. Please provide your AGE:

2. Please indicate your GENDER:

Male..... ☐

Female..... ☐

3. Please provide your current OCCUPATION/STUDY:

4. What is your FIELD of work or study?

5. What is your educational level

Undergraduate/No Degree..... ☐

Graduate Student/Primary Degree. ☐

Researcher/Advanced Degree..... ☐

Faculty/Research Staff..... ☐

6. How would you describe your proficiency with ENGLISH

Native Speaker..... ☐

Advanced..... ☐

Intermediate..... ☐

Beginner..... ☐

B.4 Entry Questionnaire

Part 2: SEARCH EXPERIENCE

Experience with Multimedia

Circle the number closest to your experience.

How often do you...	Never	Once or twice a year	Once or twice a month	Once or twice a week	Once or twice a day	More often
7. deal with photographs or images in your work, study or spare time?	1	2	3	4	5	6
8. take photographs in your work, study or spare time?	1	2	3	4	5	6
9. carry out image or video searches at home or work?	1	2	3	4	5	6

Multimedia Search Experience

10. Please indicate which online search services you use to search for MULTIMEDIA (mark AS MANY as apply)

Google (http://www.google.com).....	<input type="checkbox"/>	1
Yahoo (http://www.yahoo.com).....	<input type="checkbox"/>	2
AltaVista (http://www.altavista.com).....	<input type="checkbox"/>	3
AlltheWeb (http://www.alltheweb.com).....	<input type="checkbox"/>	4
YouTube (http://www.youtube.com).....	<input type="checkbox"/>	5
Flickr (http://www.flickr.com).....	<input type="checkbox"/>	6
Microsoft (http://www.live.com).....	<input type="checkbox"/>	7
Baidu (http://www.baidu.com).....	<input type="checkbox"/>	8
Others (please specify).....	<input type="text"/>	

11. Using the MULTIMEDIA search services you chose in question 24 is GENERALLY:

easy	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	difficult	
stressful	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	relaxing	N/A
simple	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	complex	<input type="checkbox"/>
satisfying	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	frustrating	

B.4 Entry Questionnaire

12. You find what you are searching for on any kind of MULTIMEDIA search service...

Never					Always					N/A
<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
1	2	3	4	5						

13. Please indicate which SOFTWARE TOOLS you usually use to manage your multimedia data (mark AS MANY as apply)

None (I just create directories and files on my computer).....	<input type="checkbox"/>	1
ACD See.....	<input type="checkbox"/>	2
Adobe Album/ Photoshop Elements.....	<input type="checkbox"/>	3
Picasa (Google).....	<input type="checkbox"/>	4
iPhoto (Mac).....	<input type="checkbox"/>	5
Others (please specify).....	<input type="text"/>	

14. Using the SOFTWARE TOOLS you chose in question 13 is GENERALLY:

easy	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	difficult	N/A <input type="checkbox"/>
stressful	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	relaxing	
simple	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	complex	
satisfying	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	frustrating	

15. It is easy to find/access a particular image you have saved previously on your computer.

Never					Always					N/A
<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
1	2	3	4	5						

16. Please describe your natural search strategy either online or on your computer (taking a typical search task into consideration)? (Optional)

- a) Your problem solving strategy?
- b) Is it dependent on the type of media you are seeking?
- c) In an ideal scenario, how could a system support your search strategy?

B.5 Post-Search Questionnaire for Baseline System

POST-SEARCH QUESTIONNAIRE (OB++)

To evaluate the system you have just used, we now ask you to answer some questions about it. Take into account that we are interested in knowing your opinion: answer questions freely, and consider there are no right or wrong answers. Please remember that we are evaluating the system you have just used and not you.



User ID:		System:		Task:	
----------	--	---------	--	-------	--

Please place a TICK ☒ in the square that best matches your opinion. Please answer all questions.

Part 1: TASK

In this section we ask about the search tasks you have just attempted.

1.1. The task we asked you to perform was:

unclear	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	clear
easy	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	difficult
simple	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	complex
unfamiliar	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	familiar

1.2. It was easy to formulate initial queries on these topics.

Agree		Disagree		
<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
5	4	3	2	1

1.3. The search I have just performed was.

stressful	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	relaxing
interesting	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	boring
tiring	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	restful
easy	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	difficult

1.4. I had enough time to do an effective search

Disagree		Agree		
<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
1	2	3	4	5

1.5. I believe I have succeeded in my performance of the task.

Agree		Disagree		
<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
5	4	3	2	1

B.5 Post-Search Questionnaire for Baseline System

1.6. I believe that I have found all aspects of the topic asked by the search task:

Disagree Agree

<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
1	2	3	4	5

What are the issues/problems that affected your performance?

Agree Disagree

1.7 I didn't understand the task.	1	2	3	4	5
1.8 I image collection didn't contain the image(s) I wanted.	1	2	3	4	5
1.9 The system didn't return relevant images.	1	2	3	4	5
1.10 I didn't have enough time to do an effective search.	1	2	3	4	5
1.11 I was often unsure of what action to take next.	1	2	3	4	5

Part 2: RETRIEVED IMAGES

In this section we ask you about the images you found/selected.

2.1. The images I have received through the searches were:

	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	
relevant						not relevant
inappropriate						appropriate
complete						incomplete
surprising						expected

2.2. I had an idea of which kind of images were relevant to a given topic before starting the search.

Not at all Vague Clear

<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
1	2	3	4	5

2.3. During the search I have discovered more aspects of the topic than initially anticipated.

Disagree Agree

<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
1	2	3	4	5

2.4. The image(s) I selected in the end match what I had in mind before starting the search.

Exactly Not at all

<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
5	4	3	2	1

B.5 Post-Search Questionnaire for Baseline System

2.5. I believe I have seen all possible images that satisfy my requirement.

Agree		Disagree		
<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
5	4	3	2	1

2.6. My idea of the type of images I wanted changed during performing the task.

Agree		Disagree		
<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
5	4	3	2	1

2.7. I am satisfied with the final search results I selected.

Very		Not at all		
<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
5	4	3	2	1

Part 3: SYSTEM & INTERACTION

In this section we ask you some general questions about the system you have just used.

3.1. Overall reaction to the system:

wonderful	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	terrible
satisfying	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	frustrating
easy	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	difficult
effective	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	ineffective
rigid	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	flexible
reliable	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	unreliable

3.2. When interacting with the system, I felt:

in control	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	not in control
uncomfortable	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	comfortable
confident	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	unconfident

3.3. How easy was it to USE the system?

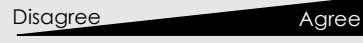
Extremely		Not at all		
<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
5	4	3	2	1

3.4. Did you find that the system response time was fast enough?

Extremely		Not at all		
<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
5	4	3	2	1

B.5 Post-Search Questionnaire for Baseline System

3.5. Browsing through the collection helped me find images I was interested in.

Disagree  Agree


☐ ☐ ☐ ☐ ☐

1 2 3 4 5

Part 4: SYSTEM SUPPORT

In this section we ask you more detailed questions about the system and your search strategy.

4.1. The system was effective for solving the task.

Agree  Disagree

☐ ☐ ☐ ☐ ☐


5 4 3 2 1

Because it helped me ...

Disagree  Agree

4.2. analyse the task	1	2	3	4	5
4.3. explore the collection.	1	2	3	4	5
4.4. find relevant images.	1	2	3	4	5
4.5. find images that I would not have otherwise considered before.	1	2	3	4	5
4.6. detect and express different aspects of the task.	1	2	3	4	5
4.7. focus my search.	1	2	3	4	5
4.8. express and illustrate your experiences.	1	2	3	4	5


4.9. The selection of relevant images from retrieved results was:

 difficult ☐ ☐ ☐ ☐ ☐ easy

effective ☐ ☐ ☐ ☐ ☐ ineffective

not useful ☐ ☐ ☐ ☐ ☐ useful

4.10. The interface supported my style of searching.

Disagree  Agree

☐ ☐ ☐ ☐ ☐

1 2 3 4 5

4.11. Which features the systems provides do you think are (would be) helpful in exploring image collections?

Quick View (automatically display images when a mouse is hovered above the image)..... ☐ 1

Transform View (The circle display which allows you to view all images in a browsing path)..... ☐ 2

Browsing Path which shows your browse history..... ☐ 3

Overall System Visualisation..... ☐ 4

Others (please specify)

B.5 Post-Search Questionnaire for Baseline System

Presentation Feature

In this section we would like to know how useful you found the presentation feature, which creates an animated slideshow.

4.12. Presentation creation (image selection and arrangement) was:

difficult	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	easy
effective	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	ineffective
not useful	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	useful

4.13. Presentation creation made you feel:

comfortable	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	uncomfortable
not in control	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	In control

4.14. An animated slideshow helps you in visualising images.

Disagree					Agree	
<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>		
1	2	3	4	5		

4.15. Do you have any other comments on the presentation feature? (optional)

- e.g. a) Did automatically generating images into a presentation panel improve exploring the image results?
 b) What could be improved?

POST-SEARCH QUESTIONNAIRE (OB++ WITH REC)

To evaluate the system you have just used, we now ask you to answer some questions about it. Take into account that we are interested in knowing your opinion: answer questions freely, and consider there are no right or wrong answers. Please remember that we are evaluating the system you have just used and not you.



User ID:		System:		Task:	
----------	--	---------	--	-------	--

Please place a TICK ☒ in the square that best matches your opinion. Please answer all questions.

Part 1: TASK

In this section we ask about the search tasks you have just attempted.

1.1. The task we asked you to perform was:

unclear	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	clear
easy	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	difficult
simple	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	complex
unfamiliar	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	familiar

1.2. It was easy to formulate initial queries on these topics.

Agree		Disagree		
<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
5	4	3	2	1

1.3. The search I have just performed was.

stressful	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	relaxing
interesting	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	boring
tiring	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	restful
easy	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	difficult

1.4. I had enough time to do an effective search

Disagree		Agree		
<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
1	2	3	4	5

1.5. I believe I have succeeded in my performance of the task.

Agree		Disagree		
<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
5	4	3	2	1

B.6 Post-Search Questionnaire for Recommender System

1.6. I believe that I have found all aspects of the topic asked by the search task:

Disagree					Agree				
<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>					
1	2	3	4	5					

What are the issues/problems that affected your performance?	Agree					Disagree				
1.6 I didn't understand the task.	1	2	3	4	5					
1.7 I image collection didn't contain the image(s) I wanted.	1	2	3	4	5					
1.8 The system didn't return relevant images.	1	2	3	4	5					
1.9 I didn't have enough time to do an effective search.	1	2	3	4	5					
1.10 I was often unsure of what action to take next.	1	2	3	4	5					

Part 2: RETRIEVED IMAGES

In this section we ask you about the images you found/selected.

2.1. The images I have received through the searches were:

relevant	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	not relevant
inappropriate	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	appropriate
complete	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	incomplete
surprising	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	expected

2.2. I had an idea of which kind of images were relevant to a given topic before starting the search.

Not at all			Vague			Clear		
<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>			
1	2	3	4	5				

2.3. During the search I have discovered more aspects of the topic than initially anticipated.

Disagree					Agree				
<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>					
1	2	3	4	5					

2.4. The image(s) I selected in the end match what I had in mind before starting the search.

Exactly			Not at all		
<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
5	4	3	2	1	

B.6 Post-Search Questionnaire for Recommender System

2.5. I believe I have seen all possible images that satisfy my requirement.

Agree		Disagree		
<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
5	4	3	2	1

2.6. My idea of the type of images I wanted changed during performing the task.

Agree		Disagree		
<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
5	4	3	2	1

2.7. I am satisfied with the final search results I selected.

Very		Not at all		
<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
5	4	3	2	1

Part 3: SYSTEM & INTERACTION

In this section we ask you some general questions about the system you have just used.

3.1. Overall reaction to the system:

wonderful	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>		terrible	
satisfying	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>		frustrating	
easy	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>		difficult	
effective	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>		ineffective	
rigid	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>		flexible	
reliable	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>		unreliable	

3.2. When interacting with the system, I felt:

in control	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>		not in control	
uncomfortable	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>		comfortable	
confident	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>		unconfident	

3.3. How easy was it to USE the system?

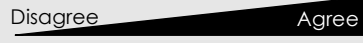
Extremely		Not at all		
<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
5	4	3	2	1

3.4. Did you find that the system response time was fast enough?

Extremely		Not at all		
<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
5	4	3	2	1

B.6 Post-Search Questionnaire for Recommender System

3.5. Browsing through the collection helped me find images I was interested in.

Disagree  Agree

☐ ☐ ☐ ☐ ☐

1 2 3 4 5

Part 4: SYSTEM SUPPORT

In this section we ask you more detailed questions about the system and your search strategy.

4.1. The system was effective for solving the task.

Agree  Disagree

☐ ☐ ☐ ☐ ☐


5 4 3 2 1

Because it helped me ...

Disagree  Agree

4.2. analyse the task	1	2	3	4	5
4.3. explore the collection.	1	2	3	4	5
4.4. find relevant images.	1	2	3	4	5
4.5. find images that I would not have otherwise considered before.	1	2	3	4	5
4.6. detect and express different aspects of the task.	1	2	3	4	5
4.7. focus my search.	1	2	3	4	5
4.8. express and illustrate your experiences.	1	2	3	4	5


4.9. The selection of relevant images from retrieved results was:

 difficult ☐ ☐ ☐ ☐ ☐ easy

effective ☐ ☐ ☐ ☐ ☐ ineffective

not useful ☐ ☐ ☐ ☐ ☐ useful

4.10. The interface supported my style of searching.

Disagree  Agree

☐ ☐ ☐ ☐ ☐

1 2 3 4 5

4.11. Which features the systems provides do you think are (would be) helpful in exploring image collections?

Quick View (automatically display images when a mouse is hovered above the image)..... ☐ 1

Transform View (The circle display which allows you to view all images in a browsing path)..... ☐ 2

Browsing Path which shows your browse history..... ☐ 3

Overall System Visualisation..... ☐ 4

Others (please specify)

B.6 Post-Search Questionnaire for Recommender System

Presentation Feature

In this section we would like to know how useful you found the presentation feature, which creates an animated slideshow.

4.12. Presentation creation (image selection and arrangement) was:

difficult	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	easy
effective	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	ineffective
not useful	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	useful

4.13. Presentation creation made you feel:

comfortable	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	uncomfortable
not in control	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	In control

4.14. An animated slideshow helps you in visualising images.

Disagree						Agree
<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	
1	2	3	4	5		

4.15. Do you have any other comments on the presentation feature? (optional)

- e.g. a) Did automatically generating images into a presentation panel improve exploring the image results?
 b) What could be improved?

B.6 Post-Search Questionnaire for Recommender System

Recommendations

In this section we would like to know how useful you found the recommendation system

4.16. The text and images recommended by the system were:

irrelevant	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	relevant
appropriate	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	inappropriate
not useful	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	useful

4.17. The recommendation made you feel:

comfortable	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	uncomfortable
not in control	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	In control

4.18. The recommendation helps me to find more images for the task.

Disagree						Agree
<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	
1	2	3	4	5		

4.19. The recommendation support me to discover more aspects of the search topic.

	Agree			Disagree	
<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
5	4	3	2	1	

4.20. I got new ideas to formulate search queries while looking at the recommended images.

Disagree						Agree
<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	
1	2	3	4	5		

4.21. Do you have any other comments on the system? (Optional)

- e.g.
- a) Did selecting images in an Ostensive Browser improve exploring the image results?
 - b) What do you think are the benefits of recommendation system based on similarity browsing?
 - c) What could be improved?

B.7 Exit Questionnaire

EXIT QUESTIONNAIRE/INTERVIEW

The aim of this experiment was to investigate the relative effectiveness of two different video search systems. Please consider the entire search experience that you just had when you respond to the following questions.

User ID:



UNIVERSITY
of
GLASGOW

Please place a TICK ☒ in the square that best matches your opinion. Please answer the questions as fully as you feel able to.

1. How different did you find the two systems from one another?

Agree

Disagree

☐

5

☐

4

☐

3

☐

2

☐

1

Which of the systems did you...	Ostensive Browsing	Ostensive Browsing with Recommendation	No difference
2 ... find BEST overall?	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
3 ... find easier to LEARN TO USE?	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
4 ... find easier to USE?	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
5 ... PERFER?	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
6 ... find changed your perception of the task?	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
7 ... find more EFFECTIVE for the tasks you performed?	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>

8 What did you LIKE about each of the systems?

System 1 (OB):

System 2 (OB+ with Recommendation):

B.7 Exit Questionnaire

3.9 What did you DISLIKE about each of the systems?

System 1 (OB):

System 2 (OB+ with Recommendation):

3.10 Additional Comments (Optional)

Appendix C

Images Features Implemented in OBP

In the Ostensive Browser Plus (OBP), the study of visual features has not been one of the main objectives. In the following sections, we briefly describe the basics of low-level signal measurements, termed features, and their particular similarity matchings, which are implemented in the OBP.

C.1 Overview of Implemented Visual Features

Table C.1: Lists of low-level features implemented in OBP.

Name	Dimensions
colour layout	12
edge histogram	80
homogeneous texture	62

C.2 Basic Image Data

An image is captured when a camera scans a scene. In digital visual capture devices, the image is represented by a two-dimensional array of individual picture elements or pixels. The density of such pixels is the resolution of a digital image. Each pixel contains the digital values that, for example, present a mixture of colours created using a certain colour model. A colour model is simply an abstract mathematical model

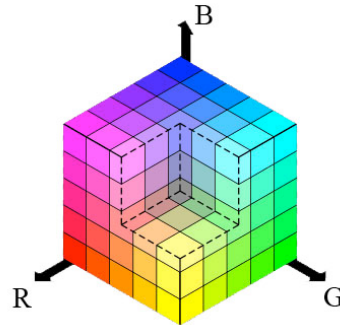


Figure C.1: RGB colour space.

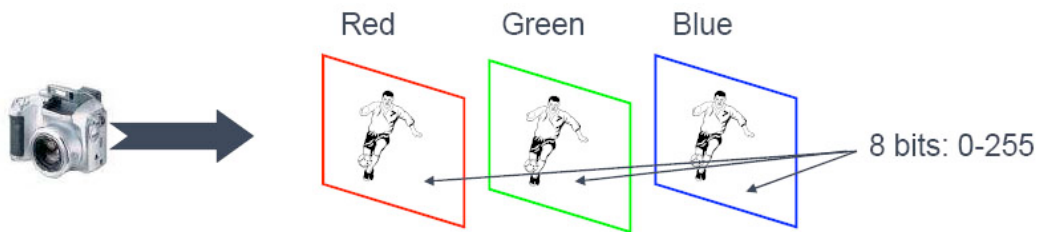


Figure C.2: Images representation using RGB colour model [O'Connor, 2009].

describing the way colours can be represented as tuples of numbers, typically as three or four values or colour components. Various colour models are commonly used in image processing such as RGB (Red, Green, Blue), HSV (Hue, Saturation, Value), CMY (Cyan, Magenta, Yellow), YCbCr (Luma, Chroma-blue, Chroma-red), and so on.

RGB is one of the most popular colour spaces, in which red, green, and blue lights are added together (additive colour model) to reproduce a broad array of colours. The mixture of these primary colours cover a large part of human colour space and thus produce a large part of human colour perception. In typical digital images based on RGB, each colour is encoded by a 8 bits integer ranging from 0 to 255. Figure C.1 shows a RGB colour space, and Figure C.2 illustrates 3 colour layers that produce a colour image.

Apart from the basic colour model used to represent images, image data can be transformed into a reduced representation set of features (referred to as feature vector)

in order to decrease the amount of data redundancy (much data, but not much information). The raw image data is often too large to be processed, in particular for indexing a large image collection and rapid retrieval. With feature extraction transforming the image data into the set of features, images can be efficiently used to perform in the desired tasks, e.g., similarity measurement. If the features extracted are carefully chosen (the features carry enough information about the image), it is expected that the features set will encode the relevant information suitable for indexing and retrieval. Examples of low-level features for image contents are colour, texture, and shape.

In recent years, MPEG-7, a standard for describing multimedia content data (e.g., image, audio, and video), has been proposed for describing and annotating audio-visual content. MPEG-7 provides the structures of metadata in a standardised way of describing in Extensible Markup Language (XML) the important concepts related to multimedia content description and management so as to facilitate searching, indexing, filtering and access. The MPEG-7 standard is aimed at supporting a broad a range of applications, devices, or computer codes. The MPEG-7 visual description included in the standard consists of basic structures and descriptors that cover the following basic visual features: colour, texture, shape, motion, localization, and face recognition. Each category consists of elementary and sophisticated descriptors. Here, we will briefly explain three low-level features implemented in OBP.

1. colour layout
2. edge histogram
3. homogeneous texture

C.3 Low-level Features

C.3.1 Colour Layout Descriptor

Colour Layout Descriptor (CLD) is very suitable for high-speed image retrieval. It is a compact and resolution-invariant descriptor. It is designed to efficiently represent the spatial distribution of colour in an image or region by clustering the image into 64 blocks and deriving the average colour of each block. These values are then transformed into a series of coefficients by performing an 8×8 Discrete Cosine Transform

(DCT). The DCT is applied to 2D array of local representative colours in YCbCr colour space, where Y is the luminance component, and Cb and Cr are the blue-difference and red-difference chrominance components. This CLD feature can be used for a wide variety of similarity-based retrieval, content filtering, and visualisation. CLD is used to measure visual similarity of images in the OBP system.

Table C.2: The representation of CLD [Salembier and Sikora, 2002].

Field	No of bits	Description
CoefficientPattern	1-2	Specifies the number of coefficients
NumberofYCcoeff	3	No of DCT coefficients for the luminance
NumberofCCcoeff	3	No of DCT coefficients for the chrominance
YCcoeff	5-6	The DCT coefficient values for luminance
CbCoeff	5-6	The DCT coefficient values for chrominance
CrCoeff	5-6	The DCT coefficient values for chrominance

Representation. The feature representation of CLD is presented in Table C.2, where the number of DCT coefficients used in CLD is variable and is presented by the CoefficientPattern field. The CoefficientPattern field can take three possible values. The first value indicates the use of six DCT coefficients for luminance and three each for chrominance, the second value indicates the use of six coefficients for both luminance and chrominance. For the third value of CoefficientPattern, the number of DCT coefficients is represented by the following NumberofYCcoeff and NumberofCCcoeff fields. The possible number of coefficients is one of 3, 6, 10, 15, 21, 28 and 64. The actual values of the coefficients are represented by the array YCcoeff, CbCoeff and CrCoeff. The lengths of each of these are either five or six bits depending on the coefficient.

Similarity matching. For matching CLDs, $A=\{DY_A, DCr_A, DCb_A\}$ and $B=\{DY_B, DCr_B, DCb_B\}$, the following distance measure can be used:

$$D(A, B) = \sqrt{\sum_i w_y(i)[DY_A(i) - DY_B(i)]^2} + \sqrt{\sum_i w_b(i)[DCb_A(i) - DCb_B(i)]^2} + \sqrt{\sum_i w_r(i)[DCr_A(i) - DCr_B(i)]^2}$$

where the parameter i represents the zigzag-scanning order of the coefficients and $w_y(i)$, $w_b(i)$, and $w_r(i)$ are the weighting factors, which should assign larger weights to the lower frequency components according to the perceptual characteristic of human vision system.

Authors/References. [Salembier and Sikora, 2002].

C.3.2 Edge Histogram Descriptor

The Edge Histogram descriptor (EHD) is an important texture descriptor for similarity search and retrieval since it reflects human image perception. Hence, it can retrieve images with similar semantic meaning. The EHD captures the spatial distribution of five types of edge categories, consisting of one non-directional edge and four namely directional edges: vertical, horizontal, 45-degree, and 135-degree, as shown in Figure C.3.

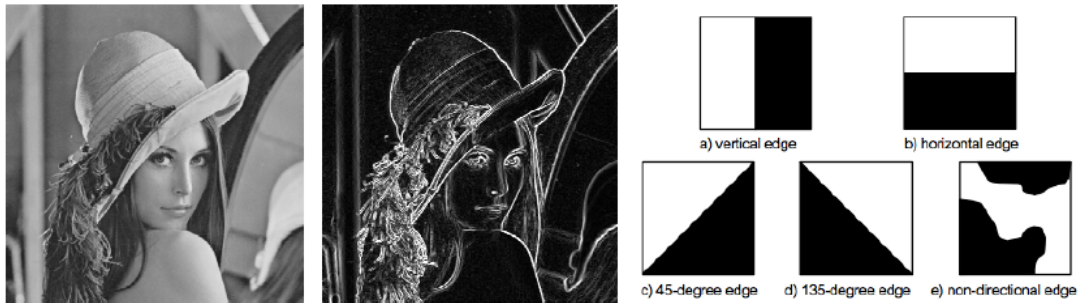


Figure C.3: Five edge types to create EHD [O'Connor, 2009].

Representation. To compute the EHD, an image is firstly divided into 16 non-overlapping sub-images (see Figure C.4). Each sub-image, called a local region, is further divided into non-overlapping image-blocks. According to MPEG-7 standard, it is suggested that the number of image-blocks around 1100 blocks in one sub-image seems to capture good directional edge feature. Each of the images-blocks is then classified into

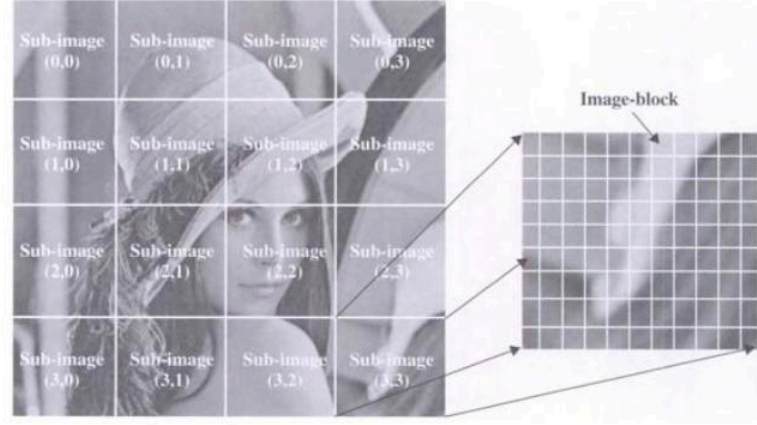


Figure C.4: Definition of sub-image and image-blocks [Salembier and Sikora, 2002].

one of the five categories so as to find the local-edge distribution represented by a histogram. Totally 80 histogram bins ($16 \text{ sub-images} \times 5 \text{ types of edge bins}$) are required to represent each edge histogram as shown in Table C.3. This simple edge histogram is called “local-edge histogram”.

The EHD primarily targets image-to-image matching (e.g., query by example or by sketch), especially for natural images with non-uniform edge distribution. In this context, the image retrieval performance can be significantly improved if the EHD is combined with other descriptors. Besides, the best retrieval performances considering this descriptor alone are obtained by using the semi-global and the global histograms generated directly from the edge histogram descriptor as well as the local ones for the matching process, as described below.

Similarity matching. Considering the local-edge histograms alone may not be sufficient enough for image matching, global edge distributions (the edge histogram of the whole images) as well as local ones are implemented. Additionally, edge distribution information for horizontal, vertical, and group of four neighbour semi-global edge distributions are required to improve the matching performance. Both global and semi-global edge histograms are estimated from the local 80 bins. The global-edge histogram is calculated by accumulating the five types of edge distributions for all sub-images. For the semi-global-edge histograms, it is estimated from the grouped sub-images, which is grouped in following ways, four groups of vertical sub-images

Table C.3: Semantics of histogram bins of the EHD [Salembier and Sikora, 2002].

\mathbf{H}_E	Semantics
$h(0)$	Relative population of vertical edges in sub-image at (0,0).
$h(1)$	Relative population of horizontal edges in sub-image at (0,0).
$h(2)$	Relative population of 45-degree edges in sub-image at (0,0).
$h(3)$	Relative population of 135-degree edges in sub-image at (0,0).
$h(4)$	Relative population of non-directional edges in sub-image at (0,0).
.	.
.	.
.	.
$h(75)$	Relative population of vertical edges in sub-image at (3,3).
$h(76)$	Relative population of horizontal edges in sub-image at (3,3).
$h(77)$	Relative population of 45-degree edges in sub-image at (3,3).
$h(78)$	Relative population of 135-degree edges in sub-image at (3,3).
$h(79)$	Relative population of non-directional edges in sub-image at (3,3).

(four columns by merging all sub-images in the same column), four groups of horizontal sub-images (four rows, similarly gathering all sub-images in the same row) and grouping of four neighbour sub-images (five groups including one group overlapping in the middle). In this case, 13 different segments are created. The corresponding edge histograms for each segment are then calculated using the local-edge histograms. After combining the local, the semi-global and the global histograms, a new histogram with 150 bins is constructed for similarity matching.

To calculate the similarity between two images in edge domain, the following distance measure using two edge histograms of images A and B is adopted, and it can be represented as:

$$D(A, B) = \sum_{i=0}^{79} |h_A(i) - h_B(i)| + 5 \times \sum_{i=0}^4 |h_A^g(i) - h_B^g(i)| + \sum_{i=0}^{64} |h_A^s(i) - h_B^s(i)|$$

where $h_A(i)$ and $h_B(i)$ are the normalised histogram bin values of image A and image B, respectively. $h_A^g(i)$ and $h_B^g(i)$ mean the normalised histogram bin values for the global-edge histograms of image A and image B, respectively. Similarly, $h_A^s(i)$ and $h_B^s(i)$

represent the histogram bin values for the semi-global-edge histograms of image A and B.

Authors/References. [Salembier and Sikora, 2002].

C.3.3 Homogeneous Texture Descriptor

The homogeneous texture descriptor (HTD) characterises the region texture using the mean energy and the energy deviation from a set of frequency channels. The 2D frequency plane is partitioned into 30 channels as shown in Figure C.5. The mean energy and its deviation are computed in each of these 30 frequency channels that corresponds to a band-limited portion of the frequency domain regarding the visual cortex in the HSV. For the details of HTD feature extraction, we refer interested readers to [Ro et al., 2001].

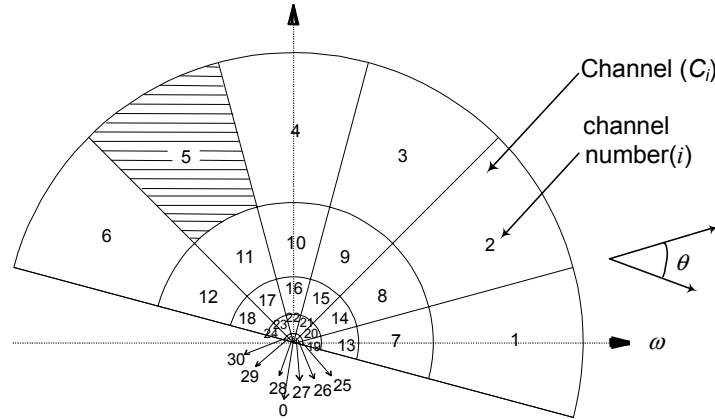


Figure C.5: 30 frequency channels used in computing the HTD [Ro et al., 2001].

Representation. The feature representation of HTD consists of 62 bins. The syntax of the HTD is as follows:

$$HTD = [f_{DC}, f_{SC}, e_1, e_2, \dots, e_{30}, d_1, d_2, \dots, d_{30}]$$

where f_{DC} and f_{SD} are the mean and standard deviation of the image, respectively, and e_i and d_i are the non-linearly scaled and quantised mean energy and energy deviation of the corresponding i -th channel in Figure C.5, respectively.

Similarity matching. To retrieve similar texture images for a query, a matching procedure should be performed. The feature of a querying image A is denoted by HTD_A while the feature of an image B in the database by HTD_B . The similarity measured by calculating the distance between the two feature vectors is as follows:

$$D(A, B) = \sum_{k=0}^{61} \left| \frac{w(k)[HTD_A(k) - HTD_B(k)]}{\alpha(k)} \right|$$

where $w(k)$ is the weighting factor of k -th descriptor value. The normalisation values $\alpha(k)$ are standard deviations of texture descriptor values. The weighting parameter $w(k)$ and the normalisation value $\alpha(k)$ are calculated in advance so that they are independent on the database. These values could be obtained a priori at the beginning of establishing the database.

Authors/References. [Ro et al., 2001; Salembier and Sikora, 2002].

Appendix D

TREC 2010 Web Diversity Track Guidelines

This appendix reports the TREC 2010 Web Diversity track evaluation guidelines, downloaded from [¹http://plg.uwaterloo.ca/~trecweb/2010.html](http://plg.uwaterloo.ca/~trecweb/2010.html). We omitted non-relevant information, e.g. the guidelines concerned with the ad-hoc retrieval task. Omissions are indicated by [...].

¹last time accessed on: 1 June 2011

TREC 2010 Web Track Guidelines

[Nick Craswell](#), Microsoft Research
[Charles Clarke](#), University of Waterloo
Ian Soboroff (NIST Contact)

[...]

For the adhoc and diversity tasks, the topic construction and judging procedures have been modified from last year. We have introduced a six-point scale for adhoc judgments. All judged runs will be fully judged according to both the adhoc and diversity criteria to some minimum depth $k \geq 10$.

[...]

Timetable

[...]

Overview

Older Web Tracks have explored specific aspects of Web retrieval, including named page finding, topic distillation, and traditional adhoc retrieval. In 2009 we introduced a new *diversity task* that combines aspects of all these older tasks. The goal of this diversity task is to return a ranked list of pages that together provide complete coverage for a query, while avoiding excessive redundancy in the result list. We continue the exploration of this task in 2010.

An analysis of last year's results indicates that the presence of spam and other low-quality pages substantially influenced the overall results. This year we are providing a [preliminary spam ranking](#) of the pages in the corpus, as an aid to groups who wish to reduce the number of low-quality pages in their results. An associated *spam task* requires groups to provide their own ranking of the corpus according to "spamminess".

In addition to the continuation of the diversity task and the introduction of the spam task, we are modifying the traditional assessment process of the *adhoc task* to incorporate multiple relevance levels, which are similar in structure to the levels used in commercial Web search. This new assessment structure includes a spam/junk level, which will assist in the evaluation of the spam task. The top two levels of the assessment structure are closely related to the homepage finding and topic distillation tasks appearing in older Web Tracks.

The adhoc and diversity tasks will share topics, which will be developed with the assistance of information extracted from the the logs of a commercial Web search engine. Topic creation and judging will attempt to reflect a mix of genuine user requirements for the topic. See below for example topics.

Document Collection

The track will again use the ClueWeb09 dataset as its document collection. The full collection consists of roughly 1 billion web pages, comprising approximately 25TB of uncompressed data (5TB compressed) in multiple languages. The dataset was crawled from the Web during January and February 2009.

Further information regarding the collection can be found on the [associated Website](#). Since it can take several weeks to obtain the dataset, we urge you to start this process as soon as you can. The collection will be shipped to you on four 1.5TB hard disks at an expected cost of US\$790 plus shipping charges.

If you are unable to work with the full dataset, we will accept runs over the smaller ClueWeb09 "Category B" dataset, but we strongly encourage you to use the full "Category A" dataset if you can. The Category B dataset represents a subset of about 50 million English-language pages. The Category B dataset can be ordered through the ClueWeb09 Web. It will be shipped to you on a single 1.0TB hard disk at an expected cost of US\$240 plus shipping charges.

Adhoc Task

[...]

Diversity Task

The diversity task is similar to the adhoc retrieval task, but differs in its judging process and evaluation measures. The goal of the diversity task is to return a ranked list of pages that together provide complete coverage for a query, while avoiding excessive redundancy in the result list. For this task, the probability of relevance of a document is conditioned on the documents that appear before it in the result list.

For the purposes of the diversity track, each topic will be structured as a representative set of subtopics, each related to a different user need. Example are provided below. Documents will be judged with respect to the subtopics. For each subtopic, NIST assessors will make a binary judgment as to whether or not the document satisfies the information need associated with the subtopic.

Topics will be fully defined by NIST in advance of topic release, but only the query field will be initially released. Detailed topics will be released only after runs have been submitted. Subtopics will be based on information extracted from the logs of a commercial search engine, and will roughly be balanced in terms of popularity. Strange and unusual interpretations and aspects will be avoided as much as possible.

Developing and validating metrics for diversity tasks continues to be a goal of the track, and we will report a number of evaluation measures that have been proposed over the past several years. These measures will include an intent aware version of the ERR measure (ERR-IA) proposed by Chapelle et al. (CIKM 2009), the α -nDCG measure proposed by Clarke et al. (SIGIR 2008), and the novelty- and rank-biased precision

(NRBP) measure proposed by Clarke et al. (ICTIR 2009). Those papers should be consulted for more information.

In all other respects, the diversity task is identical to the adhoc task. The same 50 topics will be used. Query processing must be entirely automatic. The submission format is the same. The top 10,000 documents should be submitted. You may submit up to three runs, at least one of which will be judged.

Topic Structure

The topic structure will be similar to that used for the [TREC 2009 topics](#). The topics below provide examples.

```
<topic number="6" type="ambiguous">
  <query>kcs</query>
  <description>Find information on the Kansas City Southern railroad.
</description>
  <subtopic number="1" type="nav">
    Find the homepage for the Kansas City Southern railroad.
  </subtopic>
  <subtopic number="2" type="inf">
    I'm looking for a job with the Kansas City Southern railroad.
  </subtopic>
  <subtopic number="3" type="nav">
    Find the homepage for Kanawha County Schools in West Virginia.
  </subtopic>
  <subtopic number="4" type="nav">
    Find the homepage for the Knox County School system in Tennessee.
  </subtopic>
  <subtopic number="5" type="inf">
    Find information on KCS Energy, Inc., and their merger with
    Petrohawk Energy Corporation.
  </subtopic>
</topic>

<topic number="16" type="faceted">
  <query>arizona game and fish</query>
  <description>I'm looking for information about fishing and hunting
  in Arizona.
</description>
  <subtopic number="1" type="nav">
    Take me to the Arizona Game and Fish Department homepage.
  </subtopic>
  <subtopic number="2" type="inf">
    What are the regulations for hunting and fishing in Arizona?
  </subtopic>
  <subtopic number="3" type="nav">
    I'm looking for the Arizona Fishing Report site.
  </subtopic>
  <subtopic number="4" type="inf">
    I'd like to find guides and outfitters for hunting trips in Arizona.
  </subtopic>
</topic>
```

Initial topic release will include only the *query* field.

As shown in these examples, topics are categorized as either "ambiguous" or "faceted". Ambiguous queries are those that have multiple distinct interpretations. We assume that a user interested in one interpretation would not be interested in the others. On the other

hand, facets reflect underspecified queries, with different aspects covered by the subtopics. We assume that a user interested in one aspect may still be interested in others.

Each subtopic is categorized as being either navigational ("nav") or informational ("inf"). A navigational subtopic usually has only a small number of relevant pages (often one). For these subtopics, we assume the user is seeking a page with a specific URL, such as an organization's homepage. On the other hand, an informational query may have a large number of relevant pages. For these subtopics, we assume the user is seeking information without regard to its source, provided that the source is reliable.

For the adhoc task, relevance is judged on the basis of the description field. For the diversity task, a document may not be relevant to any subtopic, even if it is relevant to the overall topic. The set of subtopics is intended to be representative, not exhaustive. We expect each topic to contain 4-10 subtopics.

Note: *We may be able to obtain probabilities indicating the relative importance of the subtopics. If so, we will include these probabilities in the topics and use them in the computation of the evaluation measures. Otherwise, we will assume subtopics are of equal importance. Further information will be posted on the mailing list in May.*

Submission Format for Adhoc and Diversity Tasks

[...]

Spam Task

[...]

Last updated: 07-Jun-2010
Date created: 29-Apr-2010
claclarke@plg.uwaterloo.ca